Seeing a Talking Face Matters to Infants, Children and Adults: Behavioural and Neurophysiological Studies

Tan Sok Hui Jessica

A thesis submitted in fulfilment of the requirements for the Degree of

Doctor of Philosophy

The MARCS Institute for Brain, Behaviour and Development

HEARing CRC

WESTERN SYDNEY UNIVERSITY

October 2020

Acknowledgements

"The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires."

—William Arthur Ward

Professor Denis Burnham, thank you for being an inspiration. Your wisdom, patience, humility, and kindness are by themselves lessons that I will never forget.

"Nothing arouses ambition so much as the trumpet clang of another's fame."

—Baltasar Gracián

Dr Marina Kalashnikova, just by being, you have shown me what I can achieve. Thank you for guiding me with your razor-sharp mind.

"Alone we can do so little; together we can do so much"

—Helen Keller

To Dr Benjawan Kasisopa, Dr Irena Lovcevic, and Dr Ruth Brookman, thank you for walking with me on this journey. This journey would have been a lot tougher without your support.

"It is not until much later, as the skin sags and the heart weakens, that children understand; their stories, and all their accomplishments, sit atop the stories of their mothers and fathers, stones upon stones, beneath the waters of their lives."

— Mitch Albom (The Five People You Meet in Heaven)

Daddo and Momma, I owe everything I have today to you.

"Sometimes our grandmas and grandpas are like grand-angels."

— Lexie Saige

To my grandparents, your unconditional love gave me the courage to march to the beat of a different drummer. You are dearly missed.

"Home isn't where you're from, it's where you find light when all grows dark."

— Pierce Brown (Golden Son)

Zi Xuan, you are my home. Thank you for never having to make me choose, for always stepping back to give way to my happiness, and for always, always believing in me.

"It is one of the blessings of old friends that you can afford to be stupid with them."

— Ralph Waldo Emerson (Emerson in His Journals)

Ahmed, Amanda, Xiaoqian, and Zhan Ning: Thank you for accepting my quirks, for always coming to my rescue, and for constantly reminding me to "be like water".

"Pets are our seat belts on the emotional roller coaster of life—they can be trusted, they keep us safe, and they sure do smooth out the ride."

— Nick Trout (Tell Me Where It Hurts: A Day of Humor, Healing and Hope in My Life As an Animal Surgeon)

What would I do without my bunnies? Seven and Summer, you make the bad days good and the good days better.

"Les grandes personnes ne comprennent jamais rien toutes seules, et c'est fatigant, pour les enfants, de toujours et toujours leur donner des explications."

[Grown-ups never understand anything by themselves, and it is tiresome for children to be always and forever explaining things to them.]

— Antoine de Saint-Exupéry (Le Petit Prince)

To my little participants and their parents, meeting you was one of the highlights in this journey. I enjoyed interacting with you, and I have learnt so much from our interactions. Thank you for adventuring with me on my quest to understand development.

Expression of Gratitude

I would also like to express my gratitude to Dr Giovanni M. Di Liberto, Dr Michael J. Crosse, and Dr Varghese Peter for introducing me to the mathematical modelling world of EEG data. To Caitlyn Hooper and the MARCS Technical Team, who helped realise the experiments I envisioned with perfect IDS stimuli, thank you.

Statement of Authentication

The work presented in this thesis, is to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

_____  _____

JESSICA

Table of Contents

List of Tables

List of Figures

Abstract

Everyday conversations typically occur face-to-face. Over and above auditory information, visual information from a speaker's face, e.g., lips, eyebrows, contributes to speech perception and comprehension. The facilitation that visual speech cues bring—termed the *visual speech benefit*—are experienced by infants, children and adults. Even so, studies on speech perception have largely focused on auditory-only speech leaving a relative paucity of research on the visual speech benefit.

Central to this thesis are the behavioural and neurophysiological manifestations of the visual speech benefit. As the visual speech benefit assumes that a listener is attending to a speaker's talking face, the investigations are conducted in relation to the possible modulating effects that gaze behaviour brings.

Three investigations were conducted. The first study examined whether visual speech information augments the accuracy with which 5-month-olds' and 4-year-olds' neural oscillations synchronise with the speech envelope. In addition, as the visual speech benefit has been previously demonstrated in adults, a group of adults were also tested to serve as a baseline. EEG and gaze data were recorded simultaneously as participants were presented with auditory-only (AO), visual-only (VO), and auditory-visual (AV) recordings of a speaker reciting short speech passages in infant-directed speech (IDS). Five-month-olds' cortical tracking was most accurate in the AV condition whereas 4-year-olds' and adults' cortical tracking accuracy did not differ between AO and AV conditions. Visual speech benefit, when quantified by the additive model criterion [i.e., AV > (A+V)], was found for 5-month-olds and adults at the frontal, occipital, and temporal scalp regions, but not for 4-year-olds.

The second study addressed whether the provision of the speaker's talking face facilitates infant word segmentation. Two groups of 7.5-month-olds were familiarised with IDS passages containing target words and then tested on their recognition of these target

words compared with non-targets over two blocks of trials. One group of infants was presented with auditory recordings paired with a still photo of the speaker's face (AO condition) while the other group was presented with auditory recordings that were paired with the corresponding video of the speaker's talking face (AV condition). Results from this study indicated that, while both groups of infants were successful at word segmentation, only infants in the AV condition demonstrated successful word segmentation performance that persisted into the second block. Interestingly, looking behaviour did not predict word segmentation performance for infants in the AV condition.

The final investigation inspected the possible relationships between neurophysiological and behavioural indices of auditory-visual speech perception in two parts. Part I examined gaze behaviour in relation to cortical tracking of the speech envelope in the first study. This revealed that 5-month-olds' relative attention to the speaker's mouth is positively correlated with their cortical tracking accuracy of the silent speech envelope, and that adults' overall attention to the screen is positively related to their cortical tracking accuracy when visual speech information is available (AV and VO conditions) in adult-directed speech (ADS), but not IDS. Part II examined cortical tracking at 5 months (Study 1) and word segmentation performance at 7.5 months (Study 2) by a subset of the infants who participated in both studies. There was no significant relationship.

Collectively, these studies demonstrate that visual speech information facilitates speech perception, and this has implications for individuals who do not have clear access to the auditory speech signal. The results, for instance the enhancement of 5-month-olds' cortical tracking by visual speech cues, and the effect of idiosyncratic differences in gaze behaviour on speech processing, expand knowledge of auditory-visual speech processing, and provide firm bases for new directions in this burgeoning and important area of research.

CHAPTER 1

Thesis Overview

1.1 The Problem of Speech Perception

The ease with which speakers of the same language understand one another paints a deceptively simple picture of speech perception. Scrutiny of the acoustic properties of speech reveals a more intricate story: formant frequencies of vowels are dependent on neighbouring consonants (Hillenbrand et al., 2001), the duration of words spoken in isolation differ considerably from the same words spoken in sentences (Pisoni, 1981), and differences in speaker characteristics such as age, gender, speech rate, accent, dialect (see Drager, 2010 for a review) and emotional tone (Mullennix et al., 2002) contribute substantially to differences in the fundamental frequency and formants values of speech sounds and phoneme category boundaries. In addition, even at the individual speaker level, no two utterances are spoken in exactly the same manner; speakers vary their phonetic realisations depending on the social situation, time of day etc., and produce novel expressions spontaneously.

Despite the highly variable and fluid nature of speech, listeners are quickly able to attend to and identify linguistically relevant and suppress linguistically irrelevant features of speech. Therein lies the problem of speech perception: how do listeners decode the highly variable acoustic signal into meaningful linguistic units and extract speakers' intended messages?

At the most basic level listeners extract the invariance of phonemes and phoneme combinations, irrespective of the variation within and between speakers across time and contexts; they extract phonetic and phonemic constancy in the sea of acoustic variation (Bradlow et al., 1999). However, over and above acoustic cues, it is now well-known that speech perception is multimodal. In particular, listeners also exploit visual cues to decode the

speech signal. Visual speech information, such as a speaker's lips and head movements, contributes to speech perception and comprehension. Numerous behavioural studies of auditory-visual speech perception by adults, children and infants have revealed better performance in auditory-visual as opposed to auditory-only conditions (e.g., Erdener & Burnham, 2013; Ross et al., 2011; Taitelbaum-Swead & Fostick, 2016). Additionally, gaze behaviour to specific facial regions, such as the eyes and mouth, may modulate the influence of visual speech information (Gurler et al., 2015). Turning to the neural level of auditory-visual speech perception, neurophysiological studies are in concert with behavioural findings. For example, adults show enhanced cortical tracking to an auditory-visual speech envelope compared to an auditory-only speech envelope (e.g., Crosse et al., 2015).

This thesis concerns auditory and auditory-visual speech perception development with a focus on infants. There is research on both these aspects of speech perception by infants, but much more is known about the former. Perceptual learning of speech patterns begins in utero such that newborns prefer their mother's voice over other female voices (DeCasper & Fifer, 1980), and also prefer stories read by their mother rather than another mother during the last six weeks of pregnancy (DeCasper & Spence, 1986). Within the first days and weeks of life, infants discriminate speech sounds (Eimas et al., 1971) and by around 4 to 6 months such discrimination becomes increasingly tuned to the linguistically significant phonemic distinctions of the specific language environment (Werker & Lalonde, 1988). At 6 months infants also recognise common words spoken by both their mother (Bergelson & Swingley, 2012) and another female (Bergelson & Swingley, 2017), and do so on the basis of semantic relations between words (Bergelson & Aslin, 2017); and later, around 9 months, infants are also able to detect and exploit transitional probabilities to learn pseudowords (Saffran et al., 1996). These findings suggest that infants begin life at least equipped with general auditory processing biases and pattern-detection strategies which they then effectively adapt to process

the language(s) that they are exposed to. Experience with the ambient language(s) in turn influences the further development of infants' speech perception. Research on infant auditory-visual speech perception augments that on auditory speech perception: visual speech information enhances phoneme discrimination (Teinonen et al., 2008), influences word processing (Weatherhead & White, 2017), and even supports infants' early recognition of their mother (Burnham, 1993). The facilitation of auditory speech perception by visual speech cues are termed as the *visual speech benefit* from here on.

The focus of this thesis is to investigate the visual speech benefit in two aspects, namely, cortical tracking of continuous speech, and segmentation of words from the speech stream.

1.2 Organisation of the Thesis

Chapter 2 begins with a review of auditory-visual speech perception in infancy, and then moves on to a consideration of the visual speech benefit in adults, children and infants. Then, as background to the empirical studies that are the focus of the thesis, the visual speech benefit in (1) cortical tracking with respect to both isolated words and continuous speech, and (2) word segmentation, is considered. Finally, the possible modulating effect of gaze behaviour on cortical tracking and word segmentation is explored.

Chapters 3, 4 and 5 concern the empirical studies of the thesis and tackle the issue of the visual speech benefit in cortical tracking, in word segmentation, and in the relationship between the two, respectively. Chapter 3 reports on a neurophysiological (EEG) study (Study 1) which examines if, like adults, 5-month-old infants and 4-year-old children demonstrate enhanced cortical tracking of the speech envelope when presented with continuous auditory-visual speech. Chapter 4 reports a behavioural study (Study 2) that investigates whether visual speech information augments infants' word segmentation from continuous speech. Given that the ability to use visual cues is contingent on gaze behavior, Study 2 also

examines the relationship between gaze behaviour and segmentation performance. Chapter 5 explores the links between the neurophysiological and behavioural indices of auditory-visual speech perception by examining (a) cortical tracking and gaze behaviour, and (b) cortical tracking and segmentation performance between 5- and 7-month-old infants. Finally, Chapter 6 discusses the results from each study, the relationships between these, and the broader implications of their results.

CHAPTER 2

Auditory-Visual Speech Perception across Development

2.1 Auditory-Visual Speech Perception in Infancy

There are at least three ways (Burnham, 1998) that an influence of visual speech information can be observed: in studies of (i) matching auditory and visual speech information, (ii) auditory-visual integration, and (iii) the perceptual benefit that results from adding visual speech information to auditory speech information. The latter, which is referred to here as the visual speech benefit is the focus of this thesis. A brief review of the other two areas are provided first followed by a section that focuses solely on the visual speech benefit.

2.1.1 Auditory-Visual Matching

Infants match auditory and visual information from the same speech event. An early study of infants' auditory-visual speech perception showed that 4.5-month-olds match auditory tokens of /a/ and /i/ to the appropriate articulating face, but not when spectral information required to identify the vowels is removed (Kuhl & Meltzoff, 1982). This finding is supported by a more recent study with 2-month-olds (Patterson & Werker, 2003), suggesting that very young infants use visual information from the orofacial area and associate phonetic information between the face and voice. The ability to match auditory and visual speech information may require very little experience as newborns are able to match auditory stimuli of sentences to corresponding point-light displays of the facial movements producing those sentences even without prior familiarisation (Guellaï et al., 2016).

Auditory-visual speech matching in older infants appears to be restricted to their native language. Ten-week-old English-learning infants were able to match visual with auditory speech when passages were presented in Greek or in English but 20-week-olds were

only able to match visual with auditory speech for the English passage (Burnham & Dodd, 1999). When presented with syllables containing Spanish homophones /b/ and /v/, Spanish-learning infants were able to match visual to auditory /ba/ and /va/ at 6 months but were unable to do so at 11 months (Pons et al., 2009). In contrast, English-learning infants were able to match visual to auditory syllables at both ages as /b/ and /v/ belong to separate phonemic categories in English (Pons et al., 2009). In support of these findings, German-learning 12-month-olds were able to match auditory stimuli to a corresponding face when the auditory stimuli consisted of German infant-directed speech (IDS) but not when the auditory stimuli consisted of French IDS (Kubicek et al., 2014). These findings are in concordance with the perceptual attunement of auditory speech perception that occurs over the first year of life (e.g., Werker & Hensch, 2015; Werker & Tees, 1984; 2005) and provide evidence that perceptual attunement occurs cross-modally; infants gain expertise in their native language auditory-visually (Danielson et al., 2017).

2.1.2 Auditory-Visual Integration

Integration involves a deeper level of auditory-visual interaction than auditory-visual matching, which could be seen as a more association-driven process. The McGurk effect is one of the most striking examples of auditory-visual integration (McGurk & MacDonald, 1976) in which auditory /ba/ dubbed onto visual /ga/ results adults' and children's illusory percept of a third sound, /da/ or /ða/ (as in 'tha'). This provides compelling evidence for the integration of auditory and visual information, as naïve perceivers unaware of the mismatch between the auditory and visual stimuli, perceive a coherent syllable in the form of a fused response. Early integration of auditory and visual speech information is shown by the fact that young infants perceive the McGurk effect (Burnham & Dodd, 2004; Desjardins & Werker, 2004; Rosenblum et al., 1997). For instance, 4.5-month-olds habituated with an auditory /ba/ dubbed onto a visual /ga/, showed a familiarity effect for the McGurk percept,

/da/ or /ða/, whereas a control group who were habituated with congruent auditory-visual /ba/ did not show the same familiarity effect (Burnham & Dodd, 2004). Desjardins and Werker (2004) and Rosenblum et al. (1997) have also found infants' perception of other variations on the basic McGurk effect, in 4.5- and 5-month-olds respectively.

Although there are results showing that the influence of visual speech increases over child age (e.g., Desjardins et al., 1997; Erdener & Burnham, 2013; Sekiyama & Burnham, 2008), studies of the McGurk effect suggest that young infants are already sensitive to visual articulatory information provided by mouth movements and that auditory and visual information are integrated in speech perception from a very young age.

2.2 Visual Speech Benefit

The McGurk effect relies upon a contrived illusion that demonstrates auditory-visual integration but does not necessarily relate to naturally occurring auditory-visual speech perception. In the real world, there is ample evidence that visual speech information *enhances* speech perception. *Visual speech benefit*, or the perceptual benefit that visual speech information brings to auditory speech perception, is the focus of this thesis. Studies of visual speech benefit are reviewed below with respect to adults, children, and infants.

2.2.1 Adults

The benefit that visual speech information provides to speech perception in the presence of auditory noise has been studied extensively. In a seminal study, Sumby and Pollack (1954) varied the speech-to-noise ratio (SNR) of auditory stimuli in auditory-only and auditory-visual presentations. The authors found that auditory-visual gain increased as SNR decreased, suggesting that the importance of visual speech information increases with increasing noise. Later studies supported this finding: speech intelligibility in cocktail-party crowd noise was higher in auditory-visual than in auditory-only condition (Schwartz et al., 2004) and the degradation of speech comprehension in the presence of multiple competing

voices was reduced when a visual display of the speaker was shown (Rudmann et al., 2003). Further, auditory-visual presentations of consonants, words, and sentences in noise resulted in faster response times and greater accuracy in identification compared to auditory-only presentations (Moradi et al., 2013). These studies indicate that visual speech information enhances speech perception in noise.

The visual speech benefit can also occur in the absence of auditory noise. It has been found that adults recognise a disyllabic word more quickly when it is paired with a video display of the articulation of the first syllable than when it is not, especially if the word has low lexical frequency (Fort et al., 2013). This raises the possibility that visual articulatory information may constrain the number of possible lexical candidates, thereby facilitating word recognition. Additionally, visual speech information can improve an individual's ability to discriminate sounds from their non-dominant language. When presented with nonsense words containing two Catalan phonemes, /ɛ/ and /e/, that are typically categorised under a single Spanish phoneme category, Spanish-dominant bilinguals (Spanish-Catalan) were able to discriminate between the two phonemes when the words were presented in the auditory-visual modality but not when the words were presented in the auditory modality (Navarra & Soto-Faraco, 2007). By comparison, Catalan-dominant bilinguals (Catalan-Spanish) were able to differentiate between the two phonemes in both modalities. These findings suggest that viewing a speaker's articulatory gestures provides an advantage over auditory-only presentations.

2.2.2 Children

Most research suggests that the visual speech benefit is evident in early childhood and that this benefit increases with age. Although 3- and 4-year-olds have been shown to be better at discriminating visually salient contrasts and recognising phonemes in auditory-visual than in auditory-only condition, adults display greater auditory-visual benefit in terms of

discriminating contrasts that are less visually salient (Lalonde & Holt, 2015). Six- to 8-year-olds are better at detecting if a speech sound is embedded in noise, discriminating between two words, and recognising if a word appeared before in auditory-visual than in auditory-only presentations (Lalonde & Holt, 2016), and, as in their previous study (Lalonde & Holt, 2015), adults displayed greater auditory-visual benefit compared to the 6- to 8-year-olds.

The same developmental improvement in auditory-visual enhancement is evident when the auditory signal is manipulated. When the auditory onsets of words were non-intact, 4- to 14-year-olds were better at identifying words presented in auditory-visual than in auditory modality, with the auditory-visual benefit increasing with age (Jerger et al., 2014). When auditory monosyllabic word tokens were masked with noise, auditory-visual gain was largest in a group of 12- to 14-year-olds, followed by a group of 8- to 11-year-olds and 5- to 7-year-olds (Ross et al., 2011). A similar pattern was observed when disyllabic words were embedded in noise; the auditory-visual gain in vowel detection scores was greater in 6- to 10-year-olds than 5- to 6-year-olds (Fort et al., 2012). When the auditory signal was degraded through noise-vocoding, 6- to 11-year-olds and adults, but not 4- to 5-year-olds, had better speech intelligibility in an auditory-visual than in an auditory-only condition (Maidment et al., 2015), suggesting that the addition of visual cues to noise-vocoded speech was of greater benefit at older ages. The lack of a visual speech benefit in 4- to 5-year-olds may have resulted from the additional demands that noise-vocoded speech placed on their speech processing capacities—younger children required greater spectral resolution in the auditory signal to understand speech as compared to older children and adults (Maidment et al., 2015).

Another possible explanation for the lack of visual speech benefit observed in 4-year-olds may involve task demands. When 4-year-olds were familiarised with words and non-words, they made greater use of visual information from a speaker's face during the test phase for non-words than words and for non-intact compared to intact auditory input when

tasked with identifying what the speaker said (Jerger, Damian, Tye-Murray, & Abdi, 2017). When consonant-vowel (CV) syllables were used instead of words (Lalonde & Holt, 2015) or sentences (Maidment et al., 2015), 4- to 5-year-olds also showed an auditory-visual gain in discriminating and identifying syllables (Jerger et al., 2017a). These findings suggest that young children make use of visual speech information to supplement the auditory signal. However, task demands may account for the variability in young children's visual speech benefit.

Taken together, studies with adults and children illustrate that visual speech information is used to augment speech perception. The developmental increase in auditory-visual gain raises the question of whether infants in their first year of life also benefit from visual speech information in language processing tasks.

2.2.3 Infants

Visual speech information has been shown to augment infants' phoneme discrimination and word segmentation. Six-month-olds discriminated between /ba/ and /da/ after watching video clips in which visual articulations of both /ba/ and /da/ were dubbed onto an auditory continuum between /ba/ and /da/ (Teinonen et al., 2008). In contrast, when the auditory continuum was paired with the visual articulation of just one syllable (either /ba/ or /da/), 6-month-olds did not perceive the /ba/-/da/ contrast. Furthermore, 7.5-month-olds have been found to be capable of segmenting words from a fluent speech stream that was blended with a background 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). Although limited, together these findings provide some evidence that infants benefit from visual speech information.

2.2.4 Summary

These studies suggest that visual speech information from a talking face benefits speech perception. The majority of studies have been conducted with adults. Findings from these studies reveal that visual speech information improves performance on speech perception tasks in quiet and in noise. Studies with children also show that children benefit from visual speech information in identifying phonemes and words that are presented in quiet or embedded in noise. The limited number of studies that have been conducted with infants suggest that they benefit from visual speech information in phoneme discrimination and word segmentation in noise. Taken together, these findings suggest that visual speech either provides information that either correlates with auditory speech or complements auditory speech when the auditory signal is degraded (see Campbell (2008) for a review).

2.3 Neural Basis of Visual Speech Benefit and Auditory-Visual Integration

Neurophysiological studies provide an additional dimension via which to examine the visual speech benefit in auditory-visual speech perception. Unlike behavioural studies, neurophysiological studies usually do not require an overt behavioural response. Further, the use of neurophysiological methods affords a glimpse into the neural processes underlying the integration of auditory and visual speech information. Studies that employ neurophysiological methods to explore auditory-visual speech perception have traditionally used event-related potentials (ERPs). While such studies provide insights into the temporal and spatial properties of cortical activity, these studies are limited by the need to use repetitive and discrete stimuli. Such isolated syllables and words do not reflect natural continuous speech. Recent neurophysiological studies have begun to investigate cortical entrainment with electroencephalogram (EEG) because this approach allows for the use of continuous speech stimuli. Each of these two types of studies are reviewed below.

2.3.1 Isolated Speech Stimuli (Traditional ERP Approach)

2.3.1.1 Adults and Children.

ERP research with 6- to 11-year-old children and adults has demonstrated that visual speech information modulates auditory processing. When 6- to 11-year-olds were presented with auditory-visual words they showed attenuated amplitude and shorter latencies of the auditory P2 ERP component as compared to auditory-only words (Knowland et al., 2014). Adults tested in the same paradigm showed attenuated amplitude and shorter latencies for auditory P2 and also for N1 in response to auditory-visual stimuli compared to auditory-only stimuli. Together these results suggest that visual speech modulation of auditory ERP components is developed, yet not *fully* developed in children (Knowland et al., 2014).

Other studies have employed either additive or subtractive methods to model the neural basis of auditory-visual integration, the visual speech benefit. The additive method entails comparing ERP responses to auditory-visual stimuli (AV) with the summation of ERP responses to auditory-only plus visual-only stimuli (A + V). Any greater response to AV than (A + V) would indicate auditory-visual integration. Using this method, Kaganovich and Schumaker (2014) revealed that peak amplitudes of auditory N1 and P2, and the latency of P2 were attenuated in 7- to 8-year-olds, 10- to 11-year-olds, and adults when they were shown auditory-visual compared to auditory-only and visual-only /ba/, /da/, and /ga/ syllables, thereby indicating auditory-visual integration. In a separate study, adult participants showed a significant shorter latency of the auditory N1/P2 response peak when presented with auditory-visual /ka/, /pa/, and /ta/ syllables as compared to auditory-only and visual-only syllables (van Wassenhove et al., 2005).

The subtractive method entails comparing auditory-only ERPs with the subtraction of visual-only ERPs from auditory-visual ERPs, i.e. [A vs (AV – V)]. One study that employed the subtractive method presented adult participants with three-syllable words or pseudowords

in which the initial and medial syllables were presented in the auditory modality with the final syllable presented either in auditory, visual, or auditory-visual modalities. The adult participants exhibited more positive ERPs in response to the final syllable of both words and pseudowords in auditory-visual modality than in auditory modality (Baart & Samuel, 2015). The subtractive method has also revealed a shorter N1 latency and a reduced P2 amplitude in response to pseudo-words, /tabi/ and /tagi/, presented in auditory-visual compared to auditory-only sine-wave speech (SWS) (Baart et al., 2014).

In general, both additive and subtractive model studies demonstrate that the integration of auditory and visual speech information is evident at the neural level and thus further suggest that visual speech cues may augment children's and adults' auditory processing.

2.3.1.2 Infants.

Most research with infants has employed behavioural methods, but neurophysiological studies provide an alternative method for examining the visual speech benefit. This method is especially useful for studying infants as an overt behavioural response is not required. The majority of the electrophysiological studies that examined 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; Kushnerenko et al., 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 illusion stimuli (auditory /ba/ dubbed onto a visual /ga/ which usually results in a "da" or "ða" 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/ and treated the integration the same as they did for congruent auditory-visual stimuli. In a study using short phrases, Hyde et al. (2011) presented 5-month-olds with an auditory recording of, "Oh, hi baby", that was either paired with a matching video of a face saying the same phrase (synchronous) or a mismatched video of a face saying a different phrase (asynchronous). 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.

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 only demonstrate the detection of a mismatch between auditory and visual stimuli; they do not show whether visual speech information augments infants' speech perception. In addition to the paucity of studies investigating visual speech benefit in infants, a major drawback of the ERP studies in general is that they necessitate the use of syllables or short phrases that are repeated multiple times in order to evoke brain responses (i.e., ERPs) which are then averaged and compared between conditions, leading to stimuli that are not entirely representative of natural fluent speech.

2.3.2 Continuous Speech Stimuli (Cortical Tracking Approach)

A growing body of work in auditory-only speech perception has begun using naturalistic speech stimuli to examine how well neural oscillations synchronise with the speech envelope. To describe the temporal alignment between neural oscillations and speech, many studies in the literature have used the terms *neural entrainment* and *cortical tracking*

interchangeably even though investigating neural entrainment, by strict definition, requires demonstrating that the cortical oscillatory activity of interest was induced by an external stimulus such as speech (see Obleser & Kayser, 2019 for a review). Here, *cortical tracking* will be used as an umbrella term to include studies that have examined neural entrainment per se and studies that have demonstrated an alignment, or synchrony, between oscillatory activity and speech. The greater the synchrony of neural oscillations with the speech envelope, the more accurate, or stronger, the cortical tracking.

2.3.2.1 Adults.

One of the earlier studies with adults presented participants with sentences that were time-compressed to between 20% and 75% of the original duration and found that neural responses in the auditory cortex tracked the speech envelope, but cortical tracking deteriorated as the speed of the sentences increased (Ahissar et al., 2001). Further, adults' sentence comprehension was strongly correlated with the strength of cortical tracking (Ahissar et al., 2001), suggesting that the synchrony of neural oscillations with the speech envelope plays a role in speech processing.

Although cortical tracking of the speech envelope is relatively stable even in noise, the strength of cortical tracking depends on attention and speech comprehension. Studies that blended auditory narrations of passages from books with spectrally-matched noise (Ding & Simon, 2013) or with the voice of a different speaker (Ding & Simon, 2012a) showed that cortical tracking of the speech envelope is relatively robust. When participants were presented with two different low-pass filtered auditory-only narrations to each ear simultaneously, neural responses in the auditory cortex were found to be synchronised with both auditory narrations, but were significantly stronger for the passage that the participants were instructed to attend to (Ding & Simon, 2012b). When noise-vocoding was used to vary the intelligibility of the speech stimuli, cortical tracking was enhanced only in the left

auditory cortex in the more intelligible condition even though a number of brain regions showed significant tracking in the unintelligible condition (Peelle et al., 2012).

2.3.2.2 Children.

Studies of the neural encoding of the speech envelope in children have mainly focused on children with dyslexia. For example, Power et al. (2016) presented noise-vocoded sentences to children with and without dyslexia to examine whether neural encoding for low frequency envelopes are impaired in children with dyslexia. They found that children with dyslexia showed significantly poorer neural encoding of low frequency amplitude information at the sentence level compared to both their chronological age-matched and their reading level-matched peers. Another study investigated phonological deficits in children with dyslexia by recording EEG data as children aged six- to 12-years with and without dyslexia listened to an auditory-only presentation of a story (Di Liberto, Peter, et al., 2018). Compared to age-matched and reading-matched controls, participants with dyslexia showed atypical low-frequency cortical tracking. In perhaps the only study that has investigated cortical tracking in typically developing children, Vander Ghinst et al. (2019) examined children's and adults' cortical tracking of speech-in-noise: compared to adults, the synchrony of neural oscillations with the speech envelope was reduced in 6- to 9-year-old children under adverse listening conditions, indicating that children's auditory system is still not as proficient as the auditory system of adults.

2.3.2.3 Infants.

Only a handful of studies have examined infants' cortical tracking of the speech envelope. Leong et al. (2017) simultaneously recorded EEG data from six- to 14-month-old infants and their mothers as they watched videos of sung nursery rhymes. Compared to adults, infants showed (a) more accurate cortical tracking of the speech envelopes at frequencies that corresponded to the rhyme and phoneme patterns of the stimuli, (b) similar

cortical tracking at frequencies that matched the syllables and prosodic stress patterns, and (c) less accurate cortical tracking at the frequency that represented slow phrasal patterns. Two other studies investigated 7-month-old infants' cortical tracking of the speech envelope through the use of ridge regression models, or temporal response functions (TRFs), that estimate the linear mapping between the speech signal and the neural response. In the first study, Kalashnikova et al. (2018) presented infants with a continuous stream of naturally-produced infant-directed speech (IDS) or adult-directed speech (ADS). Cortical tracking of the speech envelope was found for IDS but not for ADS. In the second study, Jessen et al. (2019) presented infants and adults with a cartoon movie to demonstrate that, as with adults, TRFs are a feasible means to accurately measure infants' cortical tracking of the speech envelope.

All in all, these studies reviewed above provide evidence for four main points. First, neural oscillations temporally align with the speech envelope. Second, the strength of cortical tracking is related to comprehension. Third, linguistic information, in addition to acoustic factors, drive cortical tracking. Fourth, neural oscillatory responses to the speech envelope can be found in adults, children and infants.

2.3.2.4 Cortical Tracking of Auditory-Visual Speech.

While much is known about the link between neural oscillations in the auditory cortex and the speech envelope, relatively less is known about how visual information from speech modulates cortical tracking. Even so, the few research studies conducted so far suggests that seeing a speaker's talking face augments cortical tracking in the auditory cortex to the speech envelope of the speaker (Crosse et al., 2015). Golumbic et al. (2013) presented participants with audio recordings of speakers reciting short passages that were presented either alone or paired with matching video recordings of the speaker's talking face. The researchers found that tracking of the speaker's temporal speech envelope was enhanced in the auditory-visual

compared to auditory-only condition. The augmentation of cortical tracking by visual input was more obvious when participants were presented with two auditory-only or auditory-visual speakers simultaneously and were instructed to attend to only one speaker: the capacity to selectively track a speaker's speech envelope was only evident when participants could also view the speaker's talking face (Golumbic et al., 2013). Building on these findings, Park et al. (2016) found that low-frequency brain oscillations in the left visual and motor cortices tracks the speaker's lip movements independent of the speech signal, and that cortical tracking in the left motor cortex predicts comprehension accuracy. This is consistent with behavioural findings indicating that visual speech information benefits speech perception especially in noisy environments (e.g., Sumby & Pollack, 1954), and that dynamic lip movements facilitate speech processing (Grant & Seitz, 2000). Enhanced tracking of the speaker's speech envelope when auditory-visual speech was embedded in noise was also found to be greater than the multisensory gain when auditory-visual speech was presented in quiet (Crosse et al., 2016).

Together, these studies suggest that visual speech information augments cortical tracking of the speaker's speech envelope in both quiet and in noise. Although the exact mechanism underlying the visual speech benefit in cortical tracking remains to be determined, it has been postulated that visual speech input amplifies cortical oscillatory activity generated in response to the related auditory signal (Schroeder et al., 2008). These authors hypothesised that visual speech cues (from the face and lips) can reset the phase of cortical oscillations, thereby allowing the subsequent auditory input (vocalisations) to arrive within a high-excitability phase, thus evoking an amplified response. How and why visual speech cues are able to reset the neural oscillatory phase is still unclear—future research is required.

2.3.3 Summary

In summary, neurophysiological studies provide evidence that auditory-visual interaction is evident at the neural level in adults, children and infants. The majority of the studies have used ERPs to show that visual speech information modulates auditory ERP components in adults and children. ERP studies conducted with infants have shown that infants detect the mismatch in incongruent auditory-visual stimuli. However, these ERP studies are generally limited because of the need to use discrete, isolated words or isolated syllables that are not representative of natural speech. The use of more naturalistic stimuli in the form of continuous speech is now rendered possible by studies that focus on the temporal alignment between neural oscillations and the speech envelope. These studies have shown that visual speech information enhances cortical tracking of the speaker's speech envelope in adults. *To date, no study has directly examined if visual speech information augments cortical tracking of the speech envelope in infants and children. The first goal of this thesis is to address this issue, and this will be done in Study 1 reported in Chapter 3.*

2.4 Visual Speech Benefit in Word Segmentation

Behavioural and neurophysiological studies have consistently found that visual speech information augments speech perception in adults and children. The few behavioural and neurophysiological studies with infants have shown that infants perceive the relation between auditory and visual signals, and suggest that visual speech information may benefit speech perception. In order to comprehend what a speaker is saying, one must first be able to parse words from the speaker's continuous stream of speech. It has been proposed that visual prosody, like acoustic prosody, may help with word segmentation (Mitchel & Weiss, 2014), and that infants are able to parse speech by using global prosodic features across visual and auditory modalities (Kitamura et al., 2014). As the temporal pattern of mouth movements has been found to be similar to the acoustic timescale of syllables, Chandrasekaran et al. (2009)

have speculated that movements such as the opening and closing of the mouth assist with segmentation by providing useful information regarding the onset and offset of syllables. Even though the exact mechanism underlying visual speech benefit remains unclear, it is likely that visual speech information provides additional cues that might augment the segmentation of fluent speech.

2.4.1 Word Segmentation is Facilitated by Visual Cues in Adults

Recent research with adults has made use of artificial language streams to investigate whether visual cues will enhance speech segmentation. The use of an artificial language stream removes any acoustic or semantic cues to word boundaries, leaving the listener to rely only on transitional probabilities within and between words (Saffran et al., 1996). Adults have been found to be better at identifying words that belong to an artificial language stream when the language stream to which they were familiarised was paired with a synchronous video of a speaker's talking face than when it was presented only in the auditory modality (Lusk & Mitchel, 2016; Mitchel & Weiss, 2014). This effect is specific to the congruence between the auditory and the visual speech information for adults have also segment words successfully from two artificial language streams when the auditory speech streams are paired with videos of two speakers' talking faces but not when the speech streams are paired with temporally synchronous changes in background colour or with still images of the two speakers' faces (Mitchel & Weiss, 2010). There is, nevertheless, a rider; when the two artificial speech streams were each paired with a single speaker's talking face, adults were only able to segment words from one artificial language stream (Mitchel & Weiss, 2010), suggesting that there is selective use of the *type* of visual speech information in speech perception.

Taken together, these studies provide evidence that the addition of visual cues, specifically those that come from a speaker's talking face, aid adults in segmenting

continuous speech in a newly learnt artificial language. This raises the question of whether visual speech information has the same facilitatory effect on word segmentation for infants.

2.4.2 Word Segmentation in Infants

The ability to segment continuous speech into words is essential for vocabulary development and language acquisition and is evident in infants' first year. Indeed, infants' ability to segment words from a fluent stream of speech is a strong predictor of later language skills (Von Holzen et al., 2018; Kooijman et al., 2013; Newman et al., 2006). Understanding how infants segment speech with the addition of visual information from a talker's face may provide important insights on how infants' word segmentation can be augmented.

Infant word segmentation studies have typically employed the head-turn preference procedure in which infants are familiarised with auditory stimuli consisting of either different tokens of target words or passages containing target words. Infants who were familiarised with target word tokens are tested with passages containing the familiar and unfamiliar words (e.g., Bosch et al., 2013; Houston & Jusczyk, 2000; Newman & Jusczyk, 1996; van Heugten & Johnson, 2012) while infants who were familiarised with passages are tested with tokens of familiar and unfamiliar words (e.g., Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999; Nazzi et al., 2014). In these studies, longer listening times to either the familiar word tokens or passages containing familiar words presented in the test phase indicate successful segmentation.

Infants are able to parse words from fluent speech streams by 7.5 months (Jusczyk & Aslin, 1995). Infants' segmentation ability develop over time: it is preceded by their knowledge of familiar words like their names and 'Mommy' by 6 months (Bergelson & Swingley, 2012; Bortfeld et al., 2005); and then enhanced as they begin to draw on different sources of information such as the stress patterns of their native language by 7.5 months (Curtin et al., 2005; Jusczyk, Houston, & Newsome, 1999), statistical probabilities between

syllables (Johnson & Jusczyk, 2001; Saffran et al., 1996; Thiessen & Saffran, 2003) and prosodic boundaries (Seidl & Johnson, 2006) by 8 months, phonotactic constraints by 9 months (Mattys et al., 1999), and allophonic cues by 10.5 months (Jusczyk, Hohne, & Bauman, 1999).

This capacity to identify word boundaries is influenced by several factors and becomes more stable with time. As an example, infants' performance on word segmentation tasks may be affected by differences between speakers' voices. When familiarised with repetitions of target words spoken by a female talker, 7.5-month-olds segment target words from test passages only when the test passages are spoken by a same gender (female) talker, but not by a different gender (male) talker (Houston & Jusczyk, 2000). However, when the test passages were blended with distractor passages recorded by a speaker of the opposite gender, 7.5-month-olds could segment the target words even when the signal-to-noise ratio (SNR) was 5 dB (Newman & Jusczyk, 1996), suggesting that familiarity of a voice may enhance segmentation. Another factor that has been found to facilitate word segmentation is reduplication: 9-month-olds perform better on the word segmentation task when familiarised with passages containing reduplicated words than passages containing non-reduplicated words (Ota & Skarabela, 2018), suggesting that reduplication facilitates segmentation.

Evidence for the developmental progression of infants' word segmentation abilities come from studies that compared segmentation across ages. For example, while 9-month-old infants were found to be unable to segment across different accents, 12-month-olds (Schmale et al., 2010), and 13-month-olds (Schmale & Seidl, 2009) were able to segment words from passages produced in a non-local dialectal accent and from passages produced in an unfamiliar foreign accent, respectively. Additionally, while 9-month-olds treat the presence of only a single cue to lexical stress (spectral tilt) similarly to that of multiple cues resulting in mis-segmentation, 12-month-olds show adult-like segmentation performance whereby the

presence of only a single cue is less reliable than a convergence of multiple cues to stress (Thiessen & Saffran, 2004). Further, while 10.5-month-olds do not segment bisyllabic verbs, 13.5-month-olds segment all trochaic bisyllabic verbs and iambic bisyllabic verbs that begin with consonants, and 16.5-month-olds all trochaic and iambic bisyllabic verbs regardless of word onset type (Nazzi et al., 2005).

Even though these word segmentation studies have been conducted with English-learning infants, such segmentation abilities are not confined to that language group—successful word segmentation has been found in other languages such as Catalan and Spanish (Bosch et al., 2013), Dutch (e.g., Houston et al., 2000; Kooijman et al., 2009), French (Parisian French: e.g., Mersad & Nazzi, 2012; Nazzi et al., 2014; Nishibayashi, Goyet, & Speech, 2014; Canadian French: e.g., Marquis & Shi, 2008; Polka & Sundara, 2012), German (Altvater-Mackensen & Mani, 2013; Höhle & Weissenborn, 2003) and European Portuguese (Butler & Frota, 2018). Together, the literature on word segmentation in infancy suggests that, regardless of linguistic background, infants' segmentation abilities appear around 7 months, and are further shaped by their native language becoming more robust with age.

The majority of infant word segmentation studies have focused on auditory speech, leaving a dearth of research on the role that visual speech information might play in word segmentation. To my knowledge, only one study has directly investigated whether the presentation of a speaker's talking face augments infants' word segmentation. Hollich et al. (2005) found, using auditory recordings that were always paired with a background distractor passage, 7.5-month-olds segment words from a speech stream presented with the speaker's talking face but not with a still image of the talker. In addition, infants successfully segmented words from the passages when the visual display was an oscilloscope pattern synchronised with the female's voice (similar to lip movements, amplitude deviations that were visible on the oscilloscope were greatest for syllables), suggesting that visual speech

benefit is associated with the temporal concordance of the auditory and visual information. No study has investigated whether the same facilitatory effects of visual speech information can be found even without the distractor background speech. *Accordingly, the second goal of this thesis is to examine if visual speech information augments infants' segmentation of continuous speech even in the absence of background noise and this is done in Study 2 and reported in Chapter 4.*

2.4.3 Word Segmentation and Cortical Tracking

It has been proposed that auditory-only word segmentation may occur at the prosodic (phrasal) and at the syllabic levels (Ghitza, 2017): slow modulations (below 3Hz) in temporal fluctuations of the cochlear critical-band envelopes provide overarching prosodic information related to syllables and words while faster modulations (3-20Hz) provide information related to phonetic segments within a syllable. These timescales correspond to the frequency bands of neural oscillations (Ghitza, 2011; Poeppel, 2003): prosodic features are associated with delta oscillations (<3Hz), syllables and words are associated with theta oscillations (3-9Hz), and phonetic features are associated with beta (15-30Hz) and gamma (>30Hz) oscillations. It is possible then, that the accuracy with which neural oscillations synchronise to the speech envelope reflects word segmentation performance. Evidence for this comes from very recent studies with infants (D. Choi et al., 2020) and adults (Batterink & Paller, 2017; 2019): cortical tracking accuracy during familiarisation to an artificial language was positively related to segmentation performance of the same artificial language. In addition to drawing an association between concurrent neurophysiological and behavioural markers of word segmentation, these findings, when coupled with evidence that early individual differences have enduring effects on later language abilities (e.g., Newman et al., 2015), also raise the possibility that cortical tracking may be a broad indicator of speech perception capacities, where early individual differences in cortical tracking accuracy contribute to later

performance variations observed in behavioural speech perception tasks. *Therefore, the third goal of this thesis is to explore the possible relationship between cortical tracking and later word segmentation, and is reported in Chapter 5.*

2.5 Gaze Behaviour and Its Effects on Speech Perception

Underlying the visual speech benefit is the assumption that individuals attend to a speaker's talking face. In face-to-face conversations, individuals typically seek out visual information from a speaker's face to augment speech perception and comprehension. Two key regions for conveying information are the eyes and the mouth. Generally, the eye region contains information about a speaker's emotions (Buchan et al., 2007) and intonation patterns (Lansing & McConkie, 1999), while the mouth region contains information about articulatory (Owens & Blazek, 1985) and the acoustic (Chandrasekaran et al., 2009) properties of the auditory signal.

2.5.1 Adults and Children

The facial regions on which individuals fixate are strongly dependent on the type of information that individuals seek. Adults direct greater proportions of gaze frequency and duration to the upper half of the face when identifying intonation patterns whereas they focused on the lower half of the face when identifying words (Lansing & McConkie, 1999). The type of prosodic information also plays a role in looking behaviour to the eyes and mouth regions. For example, adults look longer at the eyes than at the mouth when presented with questioning expressions whereas they attend more to the mouth when presented with focused and neutral expressions compared to questioning expressions (Simonetti et al., 2016). These findings show that looking behaviour to specific facial regions depends largely on the type of visual speech information the individual wish to acquire.

Variability in gaze fixations toward the eye and mouth regions influences speech perception. The role that individual differences in looking behaviour play in speech

perception is clearly illustrated by the between-participant variability in susceptibility to the McGurk effect. When presented with auditory /ba/ and visual /ga/ stimuli, adults who perceived the McGurk effect on at least half the trials had fixation durations predominantly localised on the talker's mouth (Gurler et al., 2015). In contrast, adults who perceived the McGurk effect on less than half the trials fixated more on the talker's eyes. In addition, individuals who looked longer to the mouth consistently reported a McGurk effect and were more likely to make use of visual speech information (Gurler et al., 2015). These findings provide evidence that individual variability in looking times to the eye and mouth regions can result in differences in auditory-visual speech perception.

Gaze behaviour changes across development and such developmental changes may influence speech perception. When presented with congruent and incongruent McGurk stimuli, 5- to 6-year-olds spend less time fixating on the mouth than do older children or adults (Irwin et al., 2017). While the relationship between gaze behavior and the McGurk effect was not directly examined, the developmental differences in fixations on the mouth suggest that the extent of visual speech influence experienced varies across ages and consequently allude to possible developmental differences in auditory-visual speech perception.

2.5.2 Infants

In addition to adults and children, infants' gaze behaviour has also been studied. Two attentional shifts in looking behavior to the eye and mouth regions have been found in infants. The first occurs between 4 and 8 months, when infants move from fixating on the eyes to the mouth of a speaker (Lewkowicz & Hansen-Tift, 2012; Pons et al., 2015). It has been posited that this shift coincides with the start of canonical babbling; infants pay more attention to interlocutors' mouth area as they begin to produce more speech sounds themselves. The second shift occurs between 10 and 12 months, when infants direct their

gaze for equivalent durations to the eyes and mouth of a speaker (Lewkowicz & Hansen-Tift, 2012; Pons et al., 2015). It is argued that 12-month-old infants no longer require as much information from the speaker's orofacial movements as they have, by this age, perceptually attuned to their native language (Werker & Tees, 1984) and have begun producing the first words of that language (Fenson et al., 1994).

Infants' looking behaviour to the eye and mouth regions suggests that they perceive the mouth region as an important conveyor of articulatory information. For example, 12-month-olds look longer at the speaker's eyes when a speaker talks in their native language but longer at the speaker's mouth when the speaker talks in a nonnative language (Kubicek et al., 2013; Pons et al., 2015), a result consistent with adults' gaze behavior in the face of an unfamiliar language (Barenholtz et al., 2016). Furthermore, 6- to 12-month-old infants were found to fixate more on the mouth when a speaker was talking than when the speaker was only smiling (Tenenbaum et al., 2013), indicating that infants' looking behaviour is associated with extracting linguistic information provided by the mouth.

Taken together, these findings provide evidence that young infants, like adults, adjust their gaze patterns when listening to a talking face. Strategic gaze-shifts to the speaker's mouth allow the registration of articulatory information, and this appears to result in enhancement of speech perception.

In general, there are three main findings in this area. First, individuals actively seek out visual speech information in speech perception tasks. Second, individual variability in gaze behaviour can account for differences in auditory-visual speech perception. Third, infants perceive that the mouth is associated with linguistic information and are able to shift their gaze to the speaker's mouth when the speaker is talking.

These studies further suggest that there may be a link between individual differences in looking behaviour and the extent of visual speech benefit gained. *Accordingly, the fourth*

goal of this thesis is to explore the relationship between individual differences in gaze behaviour and the extent of visual speech benefit experienced in (1) cortical tracking, and (2) word segmentation. As such, gaze behavior will also be recorded in Studies 1 and 2.

2.6 The Empirical Investigations

In the chapters that follow:

Chapter 3 (Study 1) examines whether visual speech information enhances the cortical tracking of the speech envelope in infants, children and adults.

Chapter 4 (Study 2) investigates whether visual speech information augments infant word segmentation, and whether infant gaze behaviour modulates their segmentation performance.

Chapter 5 (Neurophysiological-behavioural explorations) seeks to uncover any links between neurophysiological and behavioural markers of auditory-visual speech perception by examining whether gaze behaviour to the eye and mouth regions of the speaker's face modulates cortical tracking accuracy of auditory-visual speech, and whether cortical tracking accuracy at 5 months is related to later segmentation of auditory-visual speech at 7.5 months.

CHAPTER 3

Study 1: Infants' and Children's Cortical Tracking of Auditory-Visual Speech

As was seen in Chapter 2, both behavioural studies (e.g., Jerger et al., 2014; Moradi et al., 2013; Teinonen et al., 2008), and neurophysiological studies using ERP (e.g., Kaganovich & Schumaker, 2014; Knowland et al., 2014) have shown that speech perception is multimodal. In this chapter, cortical tracking (not ERP) is used to compare 5-month-old infants' and 4-year-old children's auditory-visual (AV) vs auditory-only (AO) and visual-only (VO) perception of continuous (not discrete) speech, along with a reference group of adults. The distinctive aspects of this investigation (continuous vs. discrete speech and cortical tracking vs. ERP; a measurement of AV vs. AO and VO speech perception) are considered below ahead of the description of the study.

A key limitation of the ERP studies described in Chapter 2 is the use of repetitive and short stimuli, such as isolated syllables, words or short phrases that are not entirely characteristic of natural continuous speech. A recent approach addresses this drawback by assessing cortical tracking, or the alignment between the temporal envelope of the speech input and the corresponding brain responses (e.g., Ding & Simon, 2012b; Fiedler et al., 2019; Golumbic et al., 2013; Gross et al., 2013; J. O'Sullivan et al., 2014). This approach has greater ecological validity as it allows the use of continuous rather than discrete stimuli, e.g., rather than single words, passages that more closely resemble natural speech. Accordingly, this method has been increasingly used to examine auditory-only speech perception in adults (e.g., Ding & 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). However, relatively little is known about how reliably the cortical signals track auditory-visual speech. 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, Di Liberto, & Lalor, 2016; A. O'Sullivan et al., 2019). Importantly, although there is evidence that cortical tracking can be reliably measured in children and infants, no study has investigated whether cortical tracking of auditory-visual speech is enhanced in children and infants.

In ERP studies, enhancement of speech perception in auditory-visual vs auditory-only and visual-only speech has been measured by either an additive or a subtractive modelling approach. In additive modelling the criterion for auditory-visual integration is based on the relative magnitude of neural responses to auditory-visual stimuli compared with the summation of neural responses to auditory-only and visual-only stimuli. On the other hand, in subtractive modelling the criterion for auditory-visual integration is based on the relative magnitude of neural responses to visual-only stimuli subtracted from those to auditory-visual stimuli compared with auditory-only stimuli alone. Such ERP studies with adults and children have found auditory-visual integration as indexed by attenuated amplitudes and shortened latencies of N1 and P2 components to auditory-visual speech compared to auditory-only and visual-only speech (e.g., Kaganovich & Schumaker, 2014). With respect to infants, while ERP studies have provided neural level evidence for auditory-visual integration by comparing 5-month-olds' neural responses to congruent versus incongruent auditory-visual stimuli (e.g., Hyde et al., 2011; Reynolds et al., 2013), these studies did not include auditory-only and visual-only conditions and so do not afford comparison with the modulating effect of visual information found in children and adults.

Accordingly, to allow for a direct comparison of infants' and children's neural integration of auditory and visual speech, and to do so with continuous speech, in this study 5-month-olds' and 4-year-olds' cortical tracking of the temporal speech envelope of continuous speech is examined in auditory-visual, auditory-only and visual-only conditions.

Behavioural studies have alluded to this by showing developmental differences in the extent of visual speech benefit experienced (e.g., Fort et al., 2012; Jerger et al., 2014; Maidment et al., 2015; Ross et al., 2011). The age comparisons will make clear whether the visual speech benefit at the neural level increases with age. Five-month-olds were chosen because previous behavioural (Burnham & Dodd, 2004) and ERP (e.g., Hyde et al., 2011) studies have demonstrated auditory-visual integration at 4.5 months and 5 months, respectively. Four-year-olds were chosen because findings from behavioural studies do not concur: Jerger and colleagues (2017b, 2017a) found a visual speech benefit in 4- to 5-year-olds, whereas Maidment and colleagues (2015) did not. Given that task demands may account for the variability in 4-year-olds' visual speech benefit, and that neurophysiological measures do not require an overt behavioural response, assessing 4-year-olds' cortical tracking of auditory-visual speech may elucidate the inconsistent behavioural findings. As a baseline, adults were also tested, as previous studies have shown a visual speech benefit in adults' cortical tracking of the speech envelope (e.g., Crosse et al., 2015; A. O'Sullivan et al., 2019). Gaze data were simultaneously recorded with EEG data by co-registering an eye-tracker with the EEG system to assess attention to screen during auditory-visual and visual-only trials.

Following the methods employed in adult studies of cortical tracking of auditory-visual speech (Crosse et al., 2015; Crosse, Di Liberto, & Lalor, 2016; Crosse & Lalor, 2014; A. O'Sullivan et al., 2019), cortical tracking in this study was indexed by means of ridge regression models that describe the mapping between the speech envelope and the EEG signal (in the case of VO condition, this involves mapping the EEG signal to the *unheard* speech envelope). A regression fit was then used to predict the measured EEG signals (forward modelling) and to reconstruct an estimate of the speech envelope (backward modelling) in unseen data with leave-one-out cross-validation. The quality of these predictions was taken as a measure of envelope tracking. This mathematical framework

relating ongoing speech inputs to corresponding brain responses was conducted via the use of multivariate temporal response functions (mTRFs; Crosse, Di Liberto, Bednar, & Lalor, 2016) which have been successfully implemented in infant (Jessen et al., 2019; Kalashnikova et al., 2018) and child (Di Liberto, Peter, et al., 2018) studies. Next, the additive model criterion, i.e., [AV vs. (A+V)], was used to quantify visual speech benefit; greater cortical tracking in the auditory-visual speech condition than in the algebraic sum of that in auditory-only and visual-only speech would indicate a visual speech benefit, while equivalence would suggest the absence of a visual speech benefit.

The specific hypotheses for this study are: (1) cortical tracking of the speech envelope is expected to be stronger when participants are presented with auditory-visual speech compared to auditory-only and visual-only speech conditions (AV > AO > VO), (2) a visual speech benefit, as quantified by the additive model criterion [i.e., AV > (AO +VO)], is expected, for all age groups, and (3) the extent of visual speech benefit will increase with age.

3.1 Methods

3.1.1 Participants

Five-month-olds: A final sample of 18 5-month-old infants were included (mean age = 5.49 months, SD = 0.30 months, 8 females). 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 (mean age = 4.16 years, SD = 0.14 years, 12 females). An additional 15 children were tested but excluded because five were very fidgety and did not complete the

experiment, three had excessively noisy EEG recordings and seven had insufficient gaze data.

Adults: A final sample of 18 Australian English monolingual adults aged between 18 to 56 years were included (mean 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.

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, 2012). As infant EEG studies have a typical attrition rate of 50-75% (deBoer et al., 2007), the lower bound of 15% attention was chosen in an attempt 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, 79.95% for adults.

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, parents of each participant 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.

3.1.2 Stimuli

Auditory-visual recordings of 30 short speech passages were made by a female native speaker of Australian English experienced in producing infant-directed speech (IDS; see Appendix C 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. These speech passages were adapted from recordings of IDS between mothers and their babies or from Richoz et al. (2017), and varied in durations from 8.44s to 16.35s (mean = 11.35s, SD = 1.76s). 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, visual-only and auditory-visual with the unimodal auditory and visual recordings extracted separately from the auditory-visual recordings. In the auditory-only (AO) 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 (VO) condition, the dynamic video of the speaker's talking face was presented in silence. In the auditory-visual (AV) condition, both the dynamic video and its soundtrack were played. The auditory recordings have a sampling rate of 44.1KHz 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 (10 x 3 = 30 trials). Presentation order was randomised across modalities, and 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.

3.1.3 Procedure

3.1.3.1 5-Month-Olds.

Infants sat on their mothers' laps approximately 70cm 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 1000Hz, 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 (version 16.3, 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 auditory-visual speech envelope (in Chapter 5). To this end, a Tobii X120 eye tracker was placed below the LCD screen to gather gaze fixation data. In order to run a calibration procedure and the experiment on the same computer, Presentation software was used instead of Tobii Studio. Pilot studies revealed limitations to using the Presentation software for infant calibration, specifically allowing only a limited animation duration of 1800ms at each calibration point, a duration too short to capture sufficient gaze data for calibration. To

overcome this, similar to the five-point infant calibration programmed used in Tobii Studio, gaze data were collected while small animations appeared at each of the four corners and centre of the screen. This calibration method, however, only allows for an approximation of spatial offset in gaze data.

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

3.1.3.2 4-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 & Schumaker, 2014).

Adult participants were informed that they are part of a control group for an infant and child study prior to the start of the experiment. The procedure for adults was similar to 4-year-olds, except that adult participants also participated in a second experimental session during which they were presented with similar stimuli but in adult-directed speech (ADS; details are described later in Section 3.4). The second session took place either immediately before or after this session depending on the counterbalanced order adult participants were randomly assigned to.

3.1.4 EEG Measure

3.1.4.1 Pre-Processing.

EEG data were pre-processed using EEGLAB (Delorme & Makeig, 2004), FieldTrip (Oostenveld et al., 2011), NoiseTools (http://audition.ens.fr/adc/NoiseTools/), and custom scripts in MATLAB R2019a (MathWorks, Natick, 2019). 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, Peter, et al., 2018; Folland et al., 2015; Kalashnikova et al., 2018). EEG data from the remaining 92 channels were high-pass filtered at 0.1Hz and low-pass filtered at 12Hz with Butterworth 8th order filters. As infant and child EEG recordings are noisy due to movements, artifact subspace reconstruction (ASR; Kothe & Jung, 2014) was applied to reduce 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 500ms 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 was re-referenced to the average of all channels and later downsampled to 100Hz to reduce processing time.

To investigate cortical tracking of the auditory-visual speech, stimuli were pre-processed in a manner following Jessen et al. (2019). The auditory soundtracks of each video were extracted, downsampled to 100Hz 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 4kHz. The broadband temporal envelope of each soundtrack was obtained by summing up the band-specific envelopes across all frequencies.

3.1.4.2 Data Analysis.

Cortical tracking of the speech envelope was measured, through the use of the mTRF Toolbox (Crosse, Di Liberto, Bednar, & Lalor, 2016), by mathematically modelling response functions that describe the linear mapping between the stimulus speech envelopes and the neural responses to them. The stimulus-response mapping function can either be modelled in the forward or backward direction. Forward modelling is univariate as it involves estimating a temporal response function (TRF) at every recording EEG electrode while backward modelling is multivariate as it involves mapping neural data from all recording electrodes simultaneously to generate a decoder. Model performance is evaluated by testing TRFs (forward models) on their ability to predict unseen EEG data accurately and by testing decoders (backward models) on their ability to reconstruct the speech envelope.

For this study, decoders were first generated for each condition and evaluated on their stimulus reconstruction accuracy. In comparison to TRFs, decoders are a more sensitive measure of cortical tracking because data from all recording channels are mapped simultaneously in a multivariate manner, thus providing several advantages (Crosse, Di Liberto, Bednar, & Lalor, 2016). One advantage of this approach is the increased sensitivity to signal differences between highly correlated channels since all channels are included in the model (Crosse, Di Liberto, Bednar, & Lalor, 2016). Second, stimulus features that are not explicitly encoded can be inferred from correlated encoded features (Crosse, Di Liberto, Bednar, & Lalor, 2016). Third, no pre-selection of channels is required (Crosse, Di Liberto, Bednar, & Lalor, 2016). As speech is a complex signal that is not processed as single univariate features (Yang, Wang, & Shamma, 1992), decoders provide a better quality of stimulus-response fit compared to forward TRFs. The main drawback of using decoders is that their parameters are not neurophysiologically interpretable—although it is a more sensitive measure of cortical tracking, it does not provide information beyond how accurately

the neural response encodes the stimulus features. This is where TRFs (forward models) can provide complementary information on the spatial and temporal dynamics of the neural response to the different speech types. Therefore, TRFs were computed as a second step to visualise the spatiotemporal profile of the neural response to the different stimuli (AO, VO and AV speech) presented in this study.

Response functions are typically computed for each individual participant in adult studies that modelled the stimulus-response function (e.g., Crosse et al., 2015; Crosse, Di Liberto, & Lalor, 2016; Ding & Simon, 2012)—an approach known as the *individual-subjects* approach. However, this approach requires a large (and lengthy) dataset from each participant which poses a huge challenge for infant EEG studies. Given that available infant EEG data is limited (and short), this study employed the *generic modelling* approach in which response functions (decoders and TRFs) were modelled from data of *all* participants instead as it circumvents the problem of small datasets (Di Liberto & Lalor, 2017).

The following sections describe in greater detail the backward modelling, forward modelling and generic modelling approach and the parameters used in this study.

3.1.4.2.1 Backward modelling.

The stimulus reconstruction method (for more details, see Crosse, Di Liberto, Bednar, & Lalor, 2016) was employed to investigate envelope tracking in neural responses. This involves measuring how accurately the EEG signal reconstructs the broadband speech envelopes of the stimuli. A decoder that describes the linear mapping between the stimulus speech envelopes and the neural response to them is first generated and then used to decode the speech envelope from the neural response. EEG data from all channels are mapped simultaneously in a multivariate manner to generate a single decoder for each stimulus condition (AO, VO, AV). The decoder's ability to accurately reconstruct the estimated speech envelope of the stimulus (stimulus reconstruction accuracy) is evaluated and

quantified by the correlation between the estimated and the original speech envelopes. Decoders were computed for a range of ridge parameter values ($\lambda = 10^0$ to 10^5 in steps of 0.5 on the exponent) at time lags between 0 and 600ms because no visible response was observed outside this window. Leave-one-out cross-validation (ridge regression regularization method) was used to reconstruct an estimate of each of the 30 stimulus speech envelopes per condition. The ridge parameter value that gave the highest mean correlation between the estimated speech envelope and the original stimulus speech envelope was selected.

To assess the validity of the decoders' stimulus reconstruction accuracy, decoders were also trained on shuffled envelopes that were obtained through a random shuffling of the datapoints of each broadband envelope. Reconstruction accuracy of these shuffled envelopes was expected to be similar across conditions, and significantly lower than that of the original broadband envelopes.

Visual speech benefit was examined using the additive model criterion (Stein & Meredith, 1993) as in Crosse et al. (2015). A and V decoders were constructed using neural responses to unimodal AO and VO speech conditions, and the algebraic sum of the A and V decoders (A+V) were calculated for the previously used range of ridge parameter values ($\lambda = 10^0$ to 10^5 in steps of 0.5 on the exponent) at the same time lags of 0 to 600ms. Again, the ridge parameter value that gave the highest mean correlation between the estimated and original speech envelopes was selected. The additive model (A+V) was assessed using EEG data from the AV condition. Leave-one-out cross-validation was used to quantify how accurately the (A+V) and AV decoders can reconstruct the stimulus envelopes from AV data. A significant difference between the stimulus reconstruction accuracies of these models [i.e., AV > (A+V)] was then interpreted as visual speech benefit.

3.1.4.2.2 Forward modelling.

To visualize the spatiotemporal profiles of the response functions, forward modelling was performed. The forward modelling approach involves estimating, at every channel, regression weights (temporal response function—TRF) that describes how neural responses encode the speech envelope. These TRFs represent multiple univariate mappings between the EEG signal and the stimulus, and significant non-zero weights are only observed at channels where cortical activity is related to stimulus encoding (Haufe et al., 2014). TRFs are similar to event-related potentials (ERPs) as 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 100ms refers to the impact that a change in the speech stimulus at time *t* has on the EEG at time *t* + 100ms.

TRFs were calculated for each stimulus at time lags between -200 and 1000ms before selecting a temporal region of the TRF (0-600ms) 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 with $\lambda = 100$ (chosen to maintain component amplitude) was conducted on the TRF fit to assess how well the unseen EEG data could be predicted. Prediction accuracy was quantified by calculating Pearson's *r* linear 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. As in the backward modelling approach, (A+V) response functions were computed to investigate visual speech benefit. A significantly higher prediction accuracy of AV TRFs than (A+V) TRFs was interpreted as visual speech benefit.

3.1.4.2.3 Generic modelling approach.

Adult studies that model response functions to investigate neural tracking of continuous stimuli 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 individual decoders and TRFs that are then used to model responses for the *n*th trial for each participant. This approach—*individual-subject* modelling—requires lengthy datasets for each participant that may be unattainable for the infant population. To account for the limited amount of available data from the infant sample, the *generic modelling* approach (Di Liberto & 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 reconstruct the speech envelope (backward modelling) and to predict the EEG signal (forward modelling) of the *n*th participant via leave-one-out cross-validation. The generic modelling approach has been shown to yield better results than the individual modelling approach when used with 5-minute EEG recordings from 7-month-olds and adults (Jessen et al., 2019).

3.2 Results

Separate statistical analyses were conducted for decoders and temporal response functions. As decoders have a single *r* value for each stimulus, statistical analyses in the form of Condition (AO vs. VO vs. AV) x Envelope Type (Original vs. Shuffled) repeated-measures ANOVAs were conducted to examine differences in stimulus reconstruction accuracies between conditions. As TRFs have spatiotemporal properties, estimates of global field power were computed and topographic maps of TRF weights plotted to inspect the scalp regions where TRF weights were greatest. Mean TRFs were then computed for those scalp locations identified as regions of interest (ROIs) for each condition. To evaluate TRF performance, mean prediction accuracies were derived 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 visual speech benefit and any age differences in TRF performance. Unlike stimulus reconstruction accuracies, TRF components and the respective predictive accuracy were not directly compared statistically between age groups because age-related anatomical differences may influence cortical tracking between groups independently of effects due to speech modality.

3.2.1 Decoders

Separate Condition (AO vs. VO vs. AV vs. A+V) x Envelope Type (Original vs. Shuffled) repeated-measures ANOVAs were conducted for each age group to examine whether there were any differences in stimulus reconstruction accuracy between conditions. Greenhouse-Geisser-corrected degrees of freedom are reported where sphericity was violated. Planned post-hoc analyses to investigate differences between conditions and envelope types were conducted via paired-sample *t*-tests and multiple comparisons were corrected for using the Bonferroni-Holm method. All means and standard deviations are reported in Table 1. Figure 1 depicts individual stimulus reconstruction accuracy values for each condition across age groups.

3.2.1.1 Five-Month-Olds.

As Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(27) = 56.47$, $p < .001$, the Greenhouse-Geisser correction was applied. All main effects and interactions of the 2-way repeated-measures ANOVA were significant: Condition, $F(1.41, 23.91) = 10.22$, $p < .001$, $\eta_p^2 = .38$, Envelope Type, $F(0.47, 7.97) = 67.46$, $p < .001$, $\eta_p^2 = .80$, and Condition x Envelope Type, $F(1.41, 23.91) = 15.04$, $p < .001$, $\eta_p^2 = .47$. Next, separate one-way repeated-measures ANOVAs were conducted for each Envelope Type to investigate if there were significant differences in stimulus reconstruction accuracy between conditions. The ANOVAs revealed that there were significant differences between conditions only for

the original speech envelopes, $F(3, 51) = 24.19$, $p < .001$, $\eta_p^2 = .59$, and not for the shuffled (control) envelopes, $F(3, 51) = 2.25$, $p = .09$, $\eta_p^2 = 12$.

To further examine how stimulus reconstruction accuracy differed as a function of conditions for the original speech envelopes, planned paired samples *t*-tests with adjusted alpha levels of .01 were conducted. These paired samples *t*-tests showed that stimulus reconstruction accuracy was significantly greater in the AO than the VO condition, $t(17) = 5.14$, $p < .001$, Hedges's $g = 1.21$, in AV than AO condition, $t(17) = 3.15$, $p = .006$, Hedges's $g = 0.65$, and in AV than VO condition, $t(17) = 7.56$, $p < .001$, Hedges's $g = 1.81$. However, stimulus reconstruction accuracy did not differ between AV and A+V, $t(17) = 1.05$, $p = .31$, Hedges's $g = 0.13$, indicating that there was no visual speech benefit for 5-month-olds.

3.2.1.2 Four-Year-Olds.

As Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(27) = 52.66$, $p = .002$, the Greenhouse-Geisser correction was applied. The main effects of Condition, $F(1.46, 26.20) = 12.62$, $p < .001$, $\eta_p^2 = .41$, and Envelope Type, $F(0.49, 9.73) = 8.33$, $p = .01$, $\eta_p^2 = .32$, were significant, while the Condition x Envelope Type interaction did not reach significance, $F(1.46, 26.20) = 3.09$, $p = .056$, $\eta_p^2 = .15$. As differences between conditions are of particular interest, separate one-way repeated-measures ANOVAs for each Envelope Type were conducted. These analyses indicated that there were significant differences in stimulus reconstruction accuracy between conditions for the original speech envelopes, $F(3, 54) = 9.63$, $p < .001$, $\eta_p^2 = .35$, but contrary to our predictions, this was also the case for the shuffled speech envelopes, $F(3, 54) = 4.32$, $p = .008$, $\eta_p^2 = .19$.

Planned paired-sample *t*-tests with Bonferroni-adjusted alpha levels of .05/4 were conducted on the reconstruction accuracy of the original speech envelopes. Results indicate that reconstruction accuracy was greater in AO than VO condition, $t(18) = 4.79$, $p < .001$, Hedges's $g = 1.29$, and in AV than VO condition, $t(18) = 3.70$, $p = .002$, Hedges's $g = 1.16$.

Reconstruction accuracy was not significantly different for AO and AV, $t(18) = 1.48$, $p = .16$, Hedges's $g = 0.33$, and for AV and A+V, $t(18) = -0.01$, $p = .99$, Hedges's $g = -0.002$. As the one-way ANOVA for shuffled speech was significant, paired-sample *t*-tests were also conducted for the shuffled speech envelopes. The *t*-tests showed that reconstruction accuracy of the shuffled speech envelope was greater in AV than VO condition, $t(18) = 2.81$, $p = .01$, Hedges's $g = 0.96$, and for AV than for (A+V) condition, $t(18) = 3.02$, $p = .007$, Hedges's $g = 0.69$. Differences in the reconstruction accuracy of the shuffled speech envelope were not significant between AO and AV ($t(18) = -1.06$, $p = .30$, Hedges's $g = -0.28$) and between AO and VO ($t(18) = 1.98$, $p = .06$, Hedges's $g = 0.73$). These results are unexpected. As reconstruction accuracy differed between conditions even for the shuffled speech envelopes, the results from this set of analyses must be considered with caution[1].

3.2.1.3 Adults.

The assumption of sphericity was not violated. The Condition x Envelope Type ANOVA revealed a significant main effect of Envelope Type, $F(1, 17) = 9.29$, $p = .007$, $\eta_p^2 = .35$, and a significant Condition x Envelope Type interaction, $F(3, 51) = 5.68$, $p = .002$, $\eta_p^2 = .25$. The main effect of Condition was not significant, $F(3, 51) = 2.72$, $p = .054$, $\eta_p^2 = .14$. Separate one-way repeated measures ANOVAs performed for each Envelope Type showed that there were significant differences in stimulus reconstruction accuracy between conditions only for the original speech envelopes (original speech envelopes: $F(3, 51) = 6.73$, $p < .001$, $\eta_p^2 = .28$; shuffled speech envelopes: $F(3, 51) = 1.51$, $p = .22$, $\eta_p^2 = .08$).

[1] To further examine the validity of decoders generated based on 4-year-olds' EEG data and the original speech envelopes, decoders were additionally trained on white gaussian noise. A Condition (AO vs. VO vs. AV) x Stimuli (Speech vs. Noise) repeated-measures ANOVA was conducted to further examine the validity of decoders generated based on 4-year-olds' EEG data and the original speech envelopes. The main effects of Condition and Stimuli, and the Condition x Stimuli interaction were significant (Condition: $F(3, 54) = 8.82$, $p < .001$, $\eta_p^2 = .33$; Stimuli: $F(1, 18) = 33.38$, $p < .001$, $\eta_p^2 = .65$; Condition x Stimuli: $F(3, 54) = 5.49$, $p = .002$, $\eta_p^2 = .23$). The one-way ANOVA revealed that there was no significant difference in reconstruction accuracy between conditions for decoders trained on noise ($F(3, 54) = 0.18$, $p = .91$, $\eta_p^2 = .01$).

Planned paired-sample *t*-tests (with Bonferroni-adjusted alpha levels of .05/4) performed on the reconstruction accuracy of the original speech envelopes revealed that the only significant difference in stimulus reconstruction accuracy was that the AV decoder was greater than the VO decoder, $t(17) = 4.16$, $p < .001$, Hedges's $g = 1.18$. Reconstruction accuracies did not differ between AO and VO, $t(17) = 2.22$, $p = .04$, Hedges's $g = 0.64$, AO and AV, $t(17) = -2.09$, $p = .052$, Hedges's $g = -0.54$, or between AV and A+V, $t(17) = 1.98$, $p = .06$, Hedges's $g = 0.41$.

3.2.1.4 Age Differences: Infants vs. Children vs. Adults.

An Age (5-month-olds vs. 4-year-olds vs. adults) x Condition (AO vs. VO vs. AV) mixed-measures ANOVA was conducted to investigate if reconstruction accuracy of the original speech envelopes differed between age groups. The assumption of sphericity was not violated. The two main effects and the interaction were significant: Condition, $F(2, 104) = 41.40$, $p < .001$, $\eta_p^2 = .44$, Age, $F(2, 52) = 9.93$, $p < .001$, $\eta_p^2 = .28$, and Age x Condition, $F(4, 104) = 3.93$, $p = .005$, $\eta_p^2 = .13$. Post-hoc comparisons were conducted using independent-samples *t*-tests with Bonferroni-adjusted alpha values of .05/3. These analyses revealed that 5-month-olds have greater reconstruction accuracy than adults in all conditions (AO: $t(34) = 3.64$, $p < .001$, Hedges's $g = 1.19$; VO: $t(34) = 2.97$, $p = .005$, Hedges's $g = 0.97$; AV: $t(34) = 3.86$, $p < .001$, Hedges's $g = 1.26$), and that 5-month-olds have greater reconstruction accuracy than 4-year-olds in VO ($t(35) = 3.51$, $p = .001$, Hedges's $g = 1.13$) and AV ($t(35) = 3.87$, $p < .001$, Hedges's $g = 1.25$) conditions, but not in AO condition ($t(35) = 0.65$, $p = .53$, Hedges's $g = 0.21$). No difference in stimulus reconstruction accuracy was found between 4-year-olds and adults in any condition (AO: $t(35) = 2.20$, $p = .03$, Hedges's $g = 0.71$; VO: $t(35) = -0.70$, $p = .49$, Hedges's $g = -0.23$; AV: $t(35) = -0.18$, $p = .86$, Hedges's $g = -0.06$).

To investigate whether the extent of visual speech benefit differed between age groups, a one-way between-subjects ANOVA was conducted using a difference score that

was calculated by taking the difference in reconstruction accuracy between (A+V) and AV decoders. The one-way ANOVA was not significant, $F(2, 54) = 1.04$, $p = .36$, $\eta_p^2 = .04$, suggesting that the extent of visual speech benefit did not significantly differ between age groups.

3.2.1.5 Summary

The accuracy with which the speech envelope can be reconstructed from the EEG data was measured to determine how faithfully cortical activity tracked the envelope during each condition. A comparison of the AV condition with the summation of unisensory responses (A+V) was conducted in order to examine multisensory effects.

For 5-month-olds, envelope reconstruction accuracy was greatest for AV followed by AO and VO decoders (AV > AO > VO). For 4-year-olds, envelope reconstruction accuracy was greater for both AV decoder and AO decoder than for VO decoder, with no difference between AV and AO decoders [(AV ≈ AO) > VO)]. For adults, stimulus reconstruction accuracy was greater for AV than VO decoder, with no difference between AO and AV, and AO and VO decoders [AV > VO, AV ≈ AO, AO ≈ VO]. Interestingly, no difference was found between AV and (A+V) speech for any of the three age groups, suggesting that there was no visual speech benefit in any age group.

Comparison between age groups revealed three main findings. First, envelope reconstruction accuracies of 5-month-olds were significantly greater than 4-year-olds for VO and AV speech. Second, stimulus reconstruction accuracies for 5-month-olds were significantly greater than those of adults in all conditions. Third, envelope reconstruction accuracies did not differ between 4-year-olds and adults in any condition.

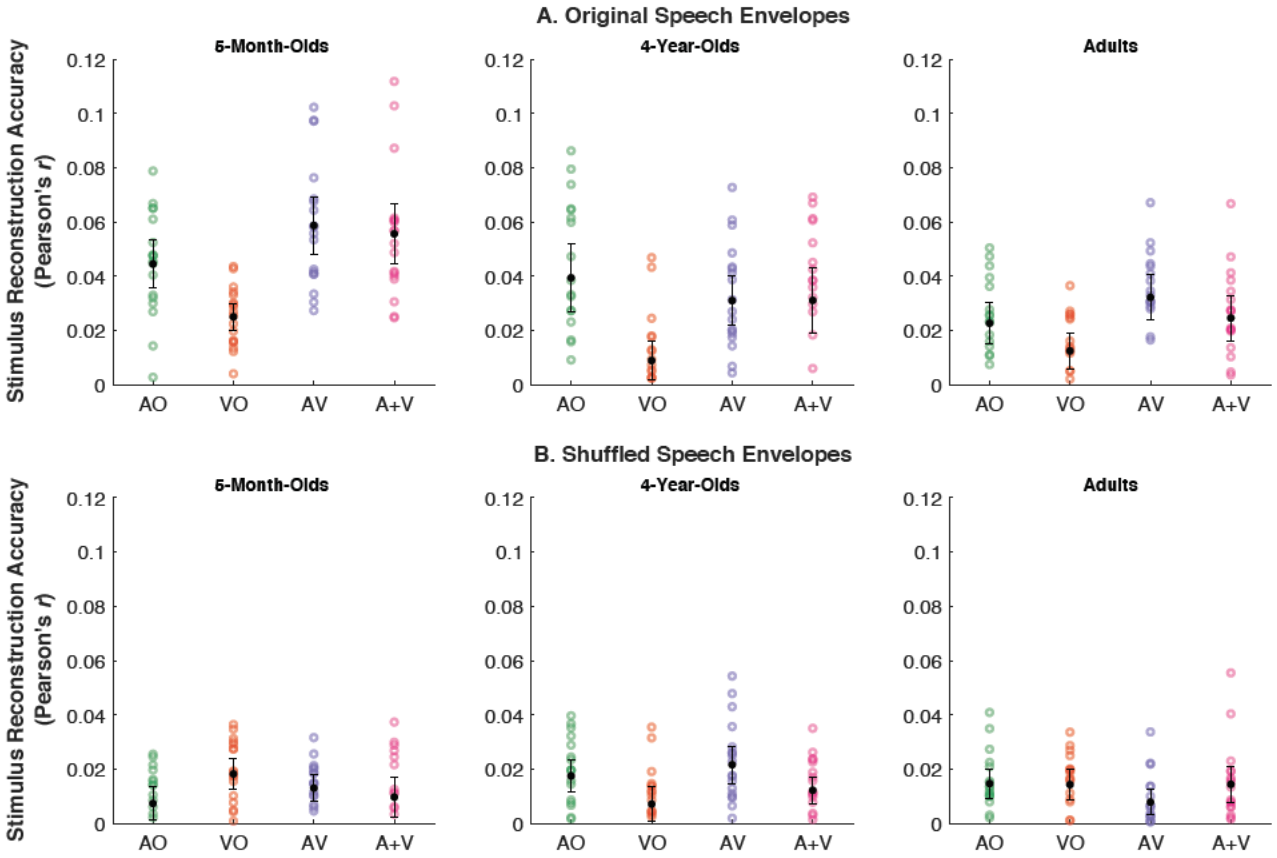
Table 1

Means (and Standard Deviations) of Stimulus Reconstruction Accuracies for All Age Groups

	AO	**VO**	**AV**	**A+V**
Original Speech Envelopes				
5-month-olds	.045 (.019)	.025 (.011)	0.059 (.022)	.056 (.024)
4-year-olds	.039 (.027)	.009 (.016)	0.031 (.020)	.031 (.026)
Adults	.023 (.016)	.012 (.014)	0.032 (.017)	.025 (.018)
Shuffled Speech Envelopes				
5-month-olds	.007 (.013)	.018 (.012)	.013 (.010)	.010 (.016)
4-year-olds	.018 (.013)	.007 (.014)	.022 (.015)	.012 (.011)
Adults	.015 (.011)	.014 (.012)	.008 (.010)	.014 (.014)

Figure 1

Stimulus Reconstruction Accuracy for All Age Groups



Individual correlations using Pearson's correlation coefficient for all age groups when decoders were trained and tested on (a) the original speech envelopes, and (b) the shuffled speech envelopes. Mean accuracies with 95% confidence intervals are shown in black.

3.2.2 Temporal Response Functions

As decoders (backward models) are not readily interpretable neurophysiologically, TRFs (forward models) provide a complementary method by which to investigate cortical tracking because the spatial and temporal information from TRFs can be examined as non-zero weights can only be observed at channels where neural activity is related to stimulus encoding (Haufe et al., 2014). Therefore, TRFs were computed at every channel for each condition, and statistical analyses were then performed on them.

First, 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 ~200-400ms for AO, AV and (A+V), but not VO (Figure 2). Topographies of TRF weights (Figures3-5) revealed that the observed components were mainly contributed by the frontal, occipital and temporal scalp regions. This is in line with previous research findings that auditory speech processing is reflected on the frontocentral region (Wunderlich et al., 2006) while visual speech processing is reflected on the parieto-occipital channels (Bernstein & Liebenthal, 2014), and the integration of auditory and visual speech information is reflected on the temporal channels (Crosse et al., 2015). 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 & Trainor, 2014) EEG studies to examine the average responses across scalp regions (Figure 6).

For visualisation purposes, mean frontal, occipital, and temporal TRFs were plotted for all age groups (Figures 7-9). Figure 10 illustrates the temporal response functions of all age groups at the three scalp ROIs. To examine the *presence* of envelope tracking, TRF

prediction accuracies at the three scalp ROIs were tested against zero. To examine 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 a visual speech benefit as quantified by the additive model [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 dependent 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.

Figure 2

Global Field Power Measured at Each Time Lag for All Ages

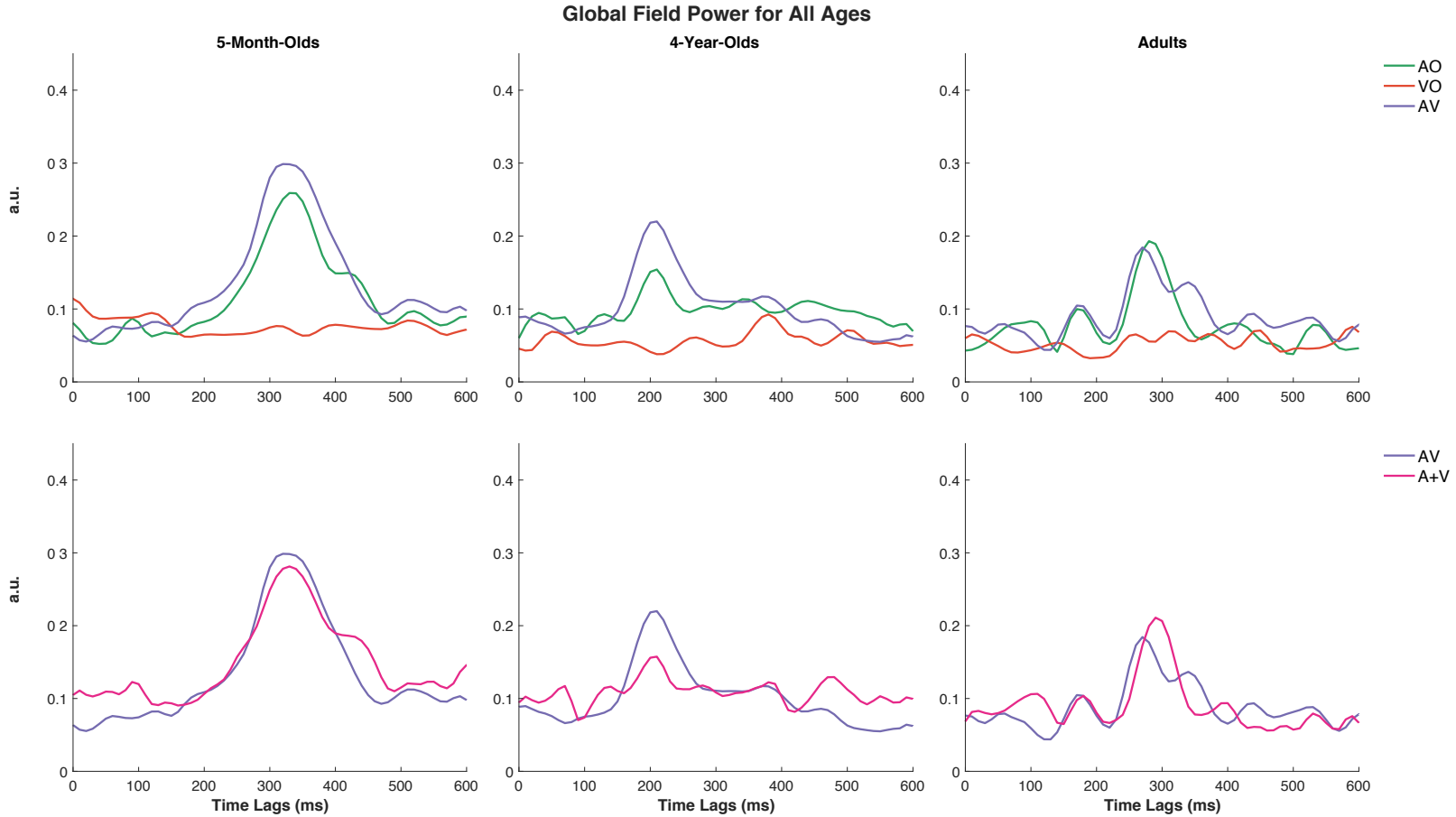
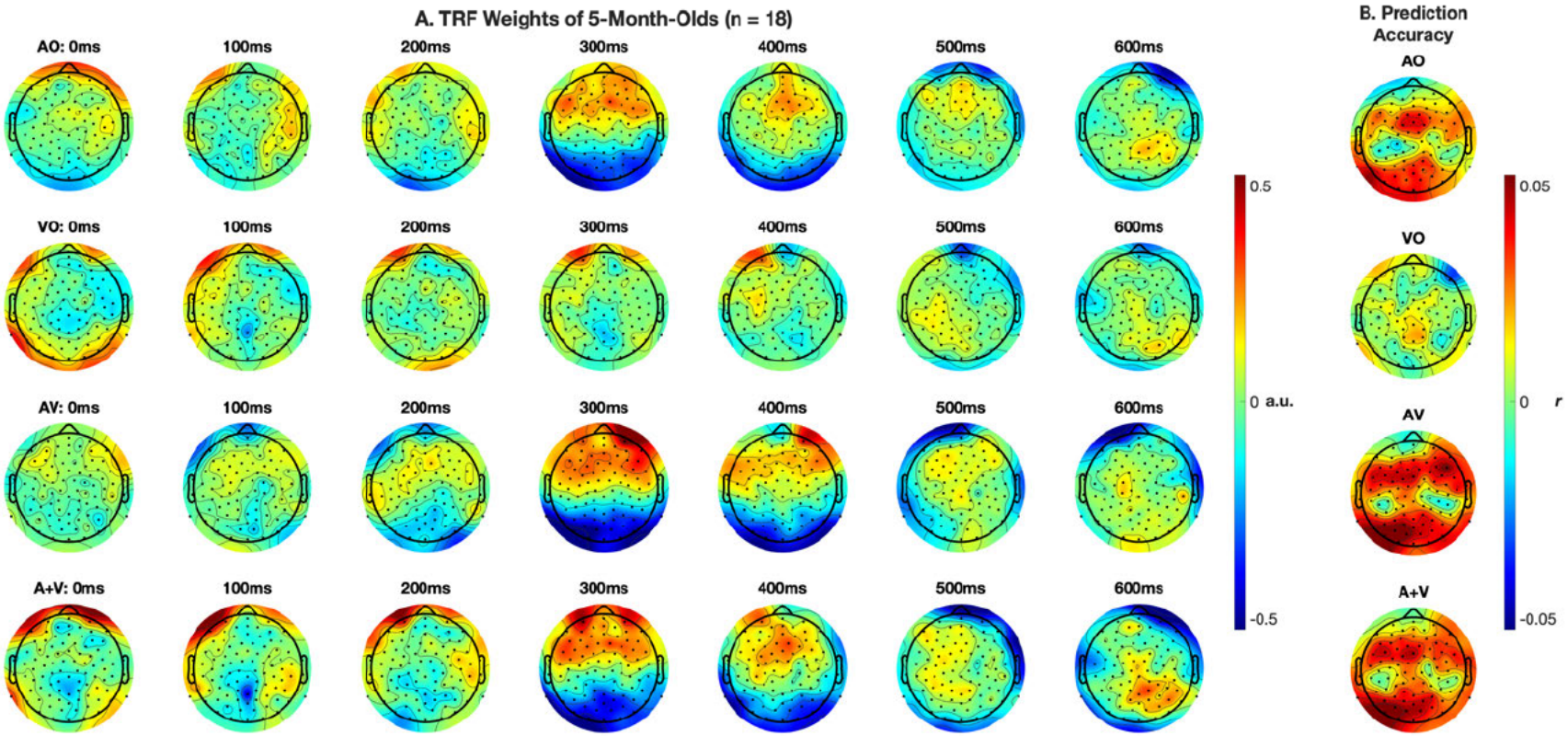


Figure 3

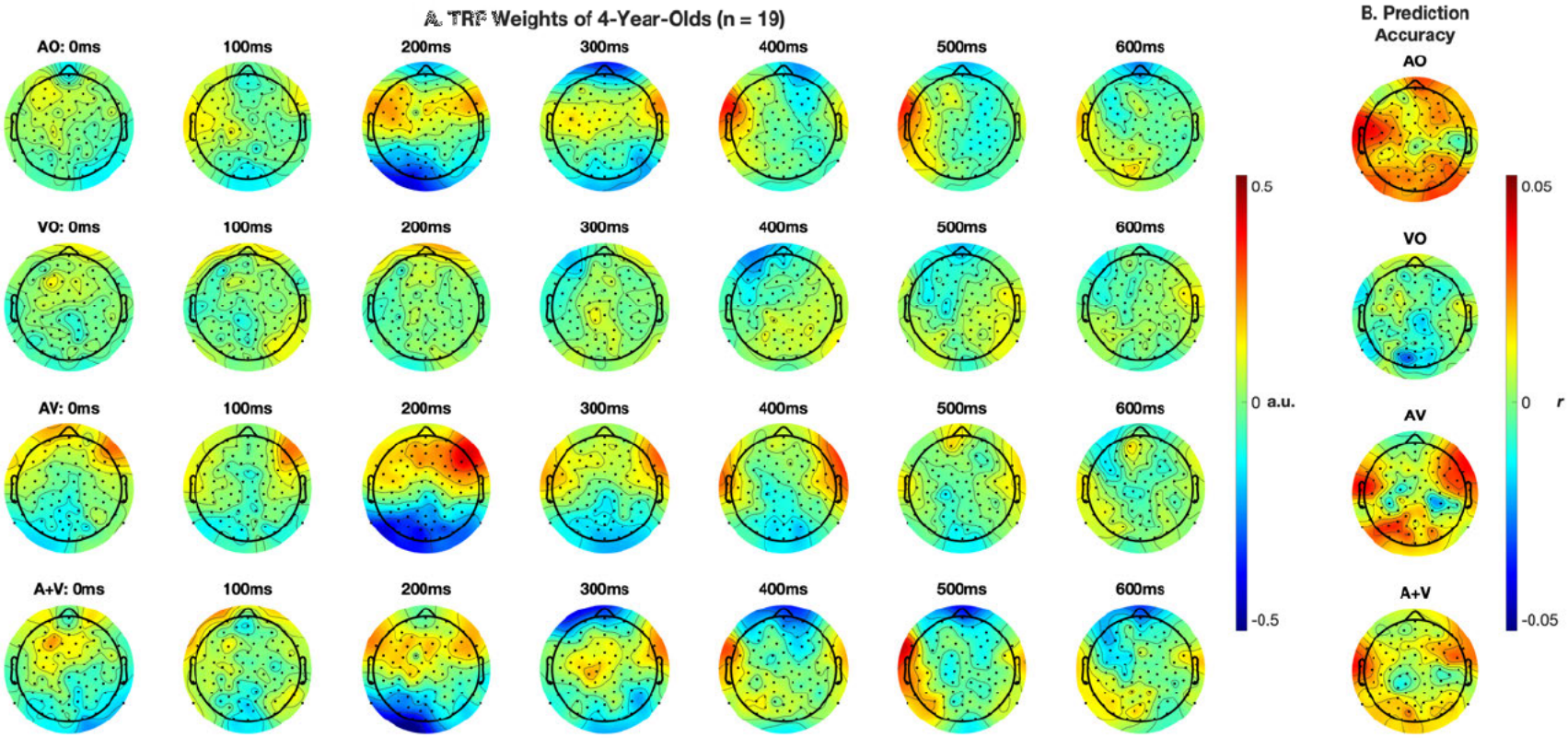
Topographies from 5-Month-Olds' Data



Topographies illustrate (a) TRF weights at time lags of 0-600ms with 100ms intervals, and (b) prediction accuracy values (Pearson's *r*) of the TRFs for each condition.

Figure 4

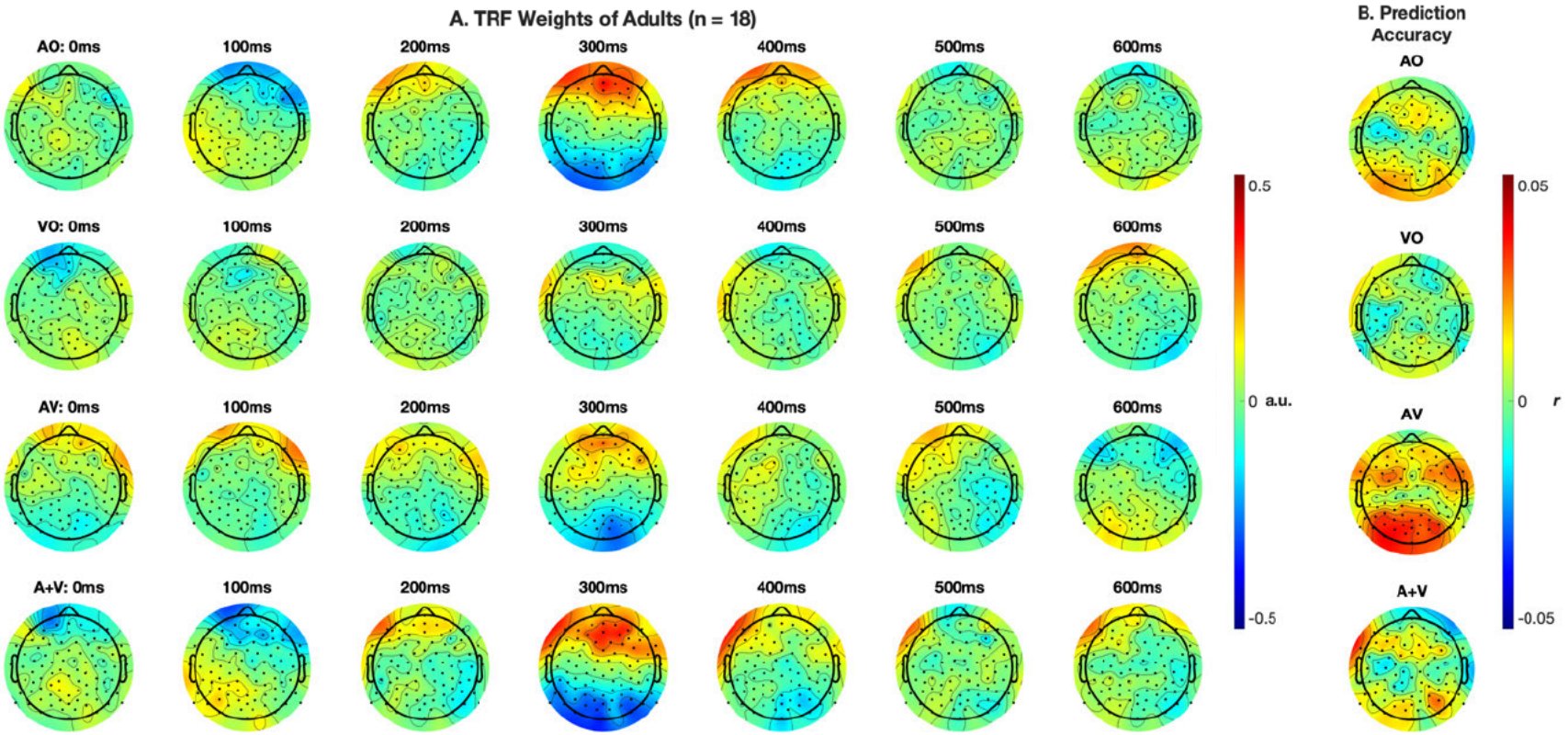
Topographies from 4-Year-Olds' Data



Topographies illustrate (a) TRF weights at time lags of 0-600ms with 100ms intervals, and (b) prediction accuracy values (Pearson's *r*) of the TRFs for each condition.

Figure 5

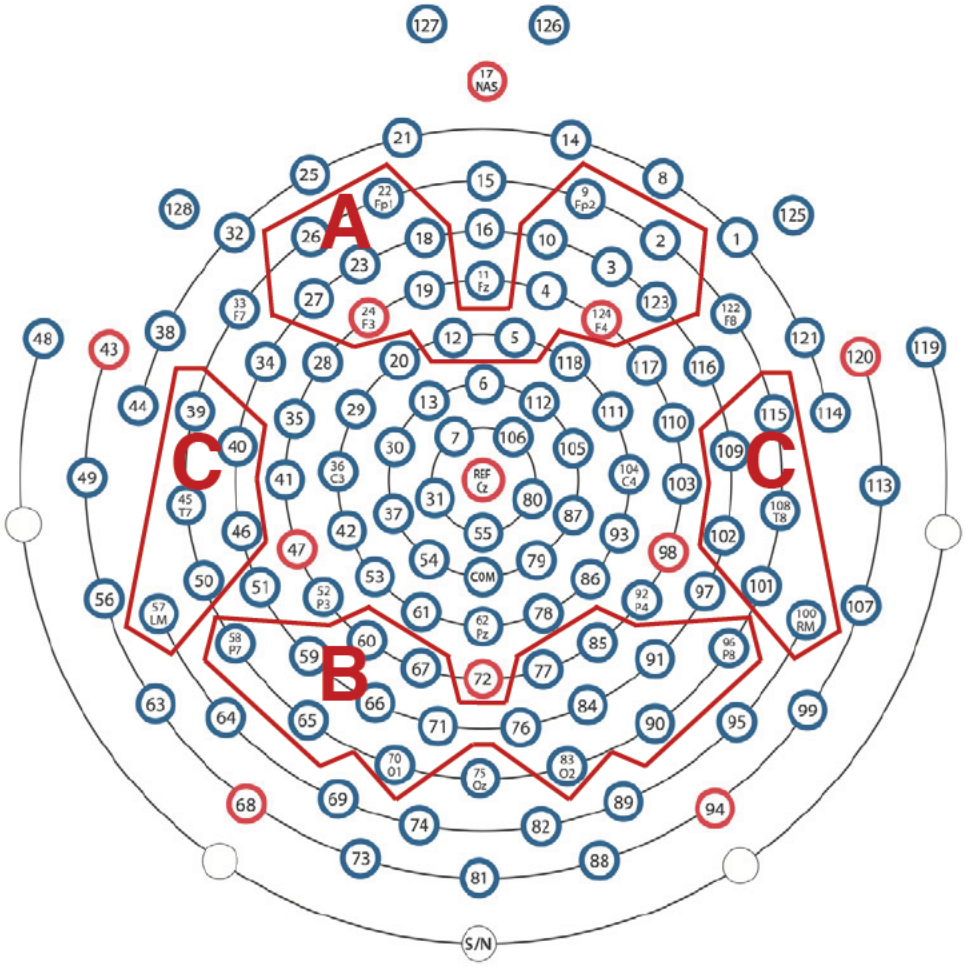
Topographies from Adults' Data



Topographies illustrate (a) TRF weights at time lags of 0-600ms with 100ms intervals, and (b) prediction accuracy values (Pearson's *r*) of the TRFs for each condition.

Figure 6

Electrode Groupings Used for Analyses. (A) Frontal Electrodes, (B) Occipital Electrodes, (C) Temporal Electrodes



3.2.2.1 Evidence of Cortical Tracking.

To examine whether participants' EEG data reflect the encoding of the speech envelope, prediction accuracies of AO, VO, AV and (A+V) TRFs were averaged across electrodes belonging to the scalp ROIs and compared against zero using one-tailed one-sample *t*-tests. All means and standard deviations of prediction accuracy for each condition and age group can be found in Table 2.

Five-month-olds: One-sample *t*-tests indicated that prediction accuracies of AO, AV, and (A+V) TRFs were significantly greater than zero (AO: $t(17) = 5.15$, $p = < .001$, Hedge's g =1.16; AV: $t(17) = 7.47$, $p < .001$, Hedge's $g = 1.68$, ; A+V: $t(17) = 7.42$, $p < .001$, Hedge's $g = 1.67$), but prediction accuracy of VO TRFs was not significantly greater than zero, $t(17) = 0.75$, $p = .23$, Hedge's $g = 0.17$.

Four-year-olds: Prediction accuracies of AO, AV, and (A+V) TRFs were significantly greater than zero (AO: $t(18) = 4.93$, $p = < .001$, Hedge's $g = 1.08$; AV: $t(18) = 3.86$, $p < .001$, Hedge's $g = 0.85$; A+V: $t(18) = 3.96$, $p < .001$, Hedge's $g = 0.87$), whereas prediction accuracy of VO TRFs was not significantly greater than zero ($t(18) = -2.13$, $p = .98$, Hedge's $g = -0.47$).

Adults: Prediction accuracies of AO, AV, and (A+V) TRFs were significantly greater than zero (AO: $t(17) = 3.49$, $p = .001$, Hedge's $g = 0.79$; AV: $t(17) = 6.11$, $p < .001$, Hedge's $g = 1.38$; A+V: $t(17) = 2.48$, $p = .012$, Hedge's $g = 0.56$), whereas prediction accuracy of VO TRFs was not significantly greater than zero, $t(17) = 0.17$, $p = .44$, Hedge's $g = 0.04$.

3.2.2.2 Difference in Strength of Cortical Tracking Between Conditions.

To examine whether the strength of envelope tracking differed between conditions, a repeated-measures one-way ANOVA was conducted for each age group with mean prediction accuracy values as the dependent variable and model type as the independent variable. The one-way ANOVAs were significant for all age groups (5-month-olds: $F(3, 68)$

= 14.95, $p < .001$, $\eta_p^2 = .40$; 4-year-olds: $F(3, 72) = 9.63$, $p < .001$, , $\eta_p^2 = .29$; adults: $F(3, 68) = 9.22$, $p < .001$, $\eta_p^2 = .29$). To inspect the differences between conditions and to identify whether there was visual speech benefit [i.e., AV > (A+V)], post hoc comparisons were subsequently performed using paired-sample *t*-tests.

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: $t(17) = 5.13$, $p < .001$, Hedge's $g = 1.42$; AO vs. AV: $t(17) = -4.07$, $p < .001$, Hedge's $g = -0.69$; AV vs. VO: $t(17) = 7.73$, $p < .001$, Hedge's $g = 2.15$). Prediction accuracy of AV TRFs was also significantly greater than (A+V) TRFs, $t(17) = 2.82$, $p = .001$, Hedge's $g = 0.16$, suggesting that 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 ($t(18) = 5.66$, $p < .001$, Hedge's $g = 1.68$) but not significantly different from the prediction accuracy of AV TRFs ($t(18) = 0.58$, $p = .57$, Hedge's $g = 0.14$). The prediction accuracy of AV TRFs was significantly greater than that of VO TRFs ($t(18) = 4.75$, $p < .001$, Hedge's $g = 1.39$), but was not significantly greater than that of (A+V) TRFs ($t(18) = 1.06$, $p = .30$, Hedge's $g = 0.21$).

Adults: Paired-sample *t*-tests showed that the prediction accuracy of AV TRFs was greatest, followed by AO, then VO TRFs (AO vs. VO: $t(17) = 4.10$, $p < .001$, Hedge's $g = 0.78$; AO vs. AV: $t(17) = -3.85$, $p = .001$, Hedge's $g = -0.88$; AV vs. VO: $t(17) = 7.36$, $p < .001$, Hedge's $g = 1.57$). Prediction accuracy of AV TRFs was also significantly greater than (A+V) TRFs ($t(17) = 5.01$, $p < .001$, Hedge's $g = 1.06$), suggesting that visual speech benefit was present at the scalp ROIs.

3.2.2.3 Summary.

AO, VO and AV TRFs as well as the sum of unimodal TRFs (A+V) were computed to examine the spatial and temporal profile of the observed findings from backward modelling. Estimates of global field power for each age group displayed clear components for AO, AV and (A+V) TRFs at ~200-400ms, whereas there was no such observation was made for VO TRFs. Next, visual inspection of topographies of TRF weights revealed that TRF weights were greatest at frontal, occipital and temporal regions. As these regions have also been previously found to be implicated in auditory and visual speech processing, subsequent statistical analyses were focused on the prediction accuracies of the TRFs generated at these regions.

To address whether EEG data at these regions reliably encoded the speech envelope, prediction accuracies were tested against zero. Across the three age groups, prediction accuracies of AO, AV and (A+V) TRFs were significantly greater than zero, whereas the prediction accuracy of VO TRFs was not. This provides evidence that neural responses at the frontal, occipital and temporal regions were reliably tracking the speech envelope. Next, to examine whether the strength of cortical tracking differed as a function of presentation modality, prediction accuracies were compared between conditions. The prediction accuracy of AV TRFs was strongest followed by that of AO TRFs, and that of VO TRFs for 5-month-olds and adults. By comparison, the prediction accuracies of AO and AV TRFs for 4-year-olds were not significantly different from each other but were significantly greater than that of VO TRFs.

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 (.018)	.001 (.008)	.035 (.019)	.032 (.018)
4-year-olds	.020 (.018)	-.005 (.011)	.018 (.020)	.014(.015)
Adults	.009 (.011)	.0004 (.011)	.022 (.015)	.007 (.012)

Figure 7

Temporal Response Functions for Five-Month-Olds at Frontal, Occipital, and Temporal Scalp Locations

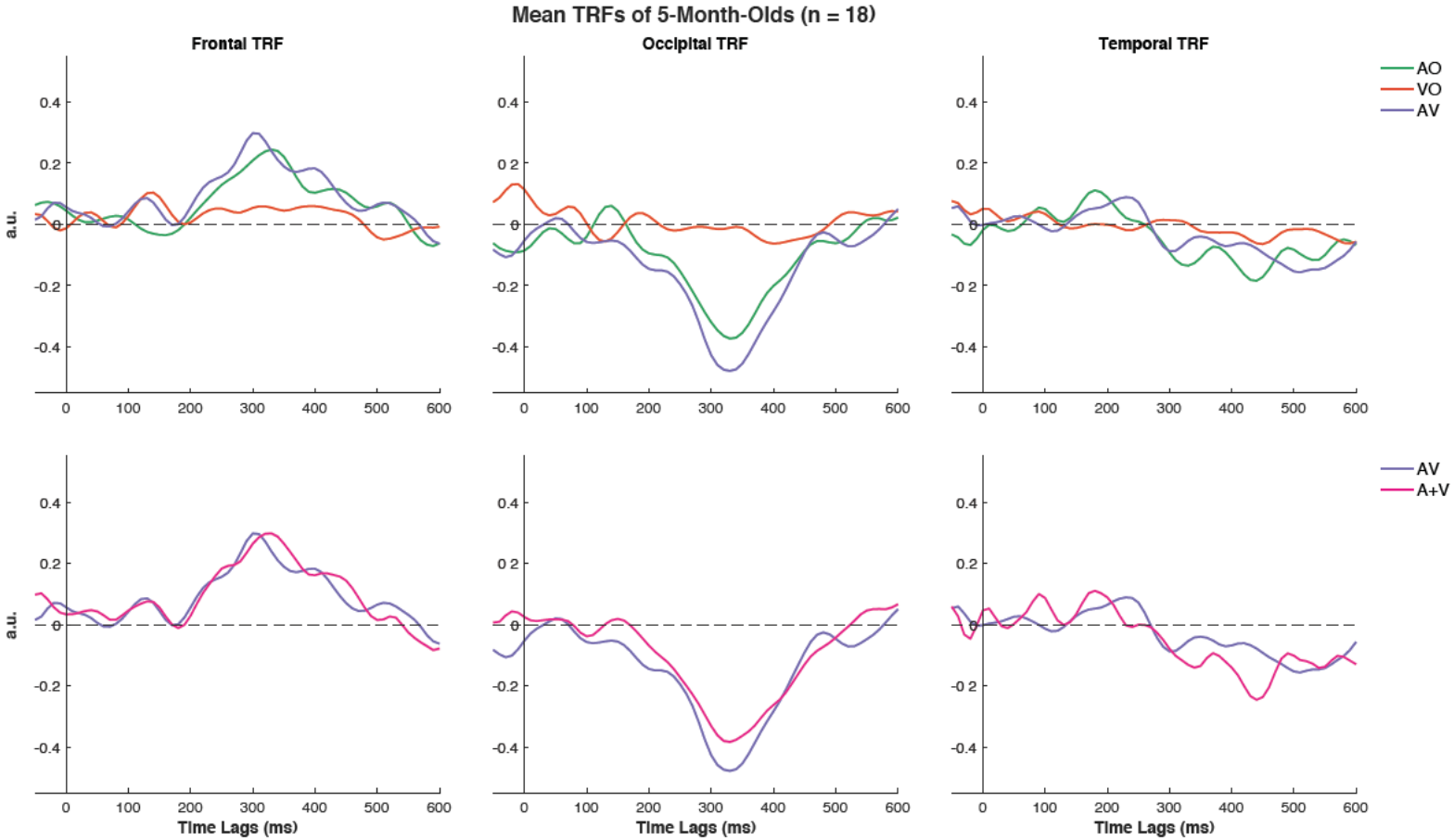


Figure 8

Temporal Response Functions for Four-Year-Olds at Frontal, Occipital, and Temporal Scalp Locations

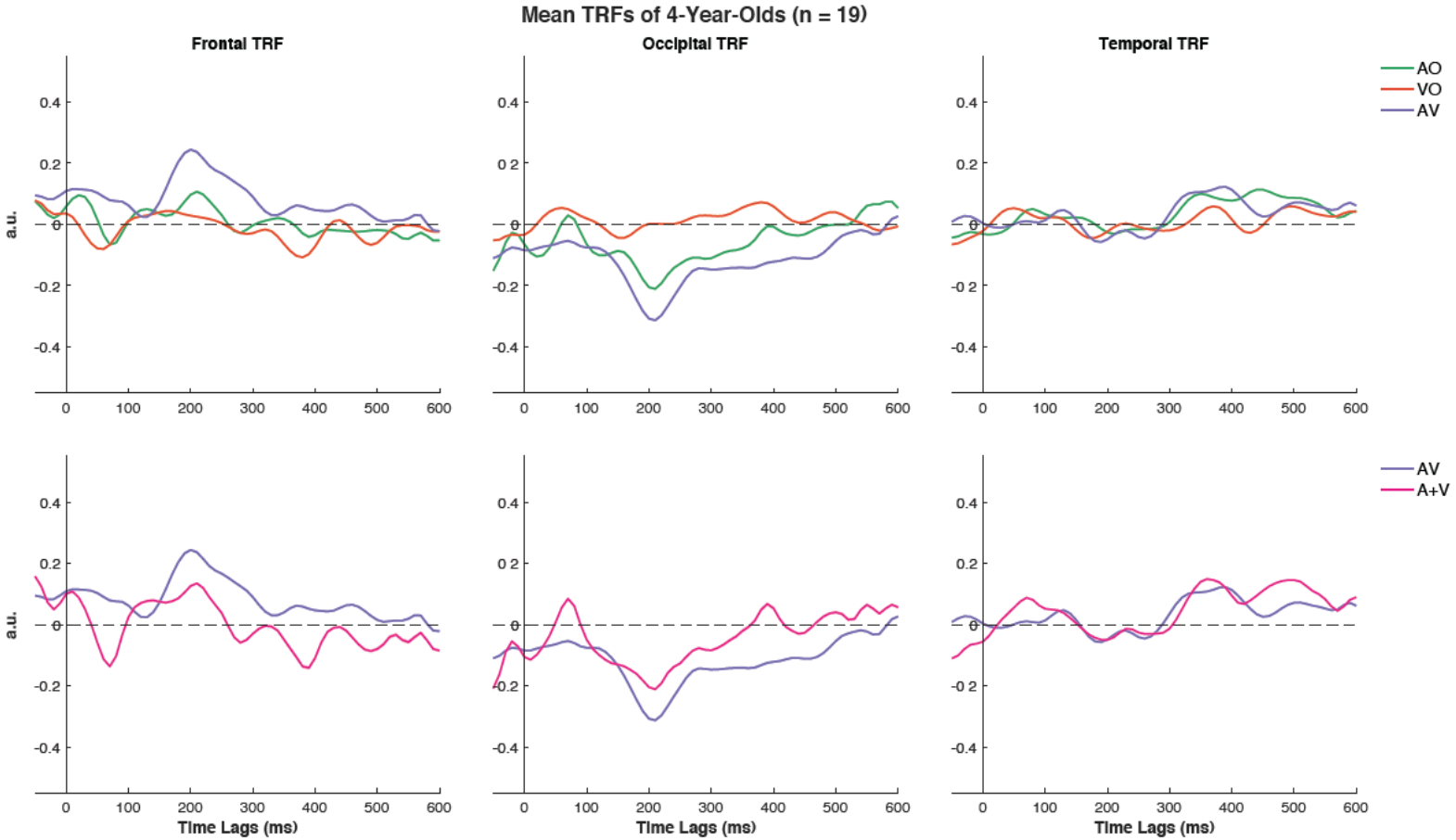


Figure 9

Temporal Response Functions for Adults at Frontal, Occipital, and Temporal Scalp Regions and Global Field Power

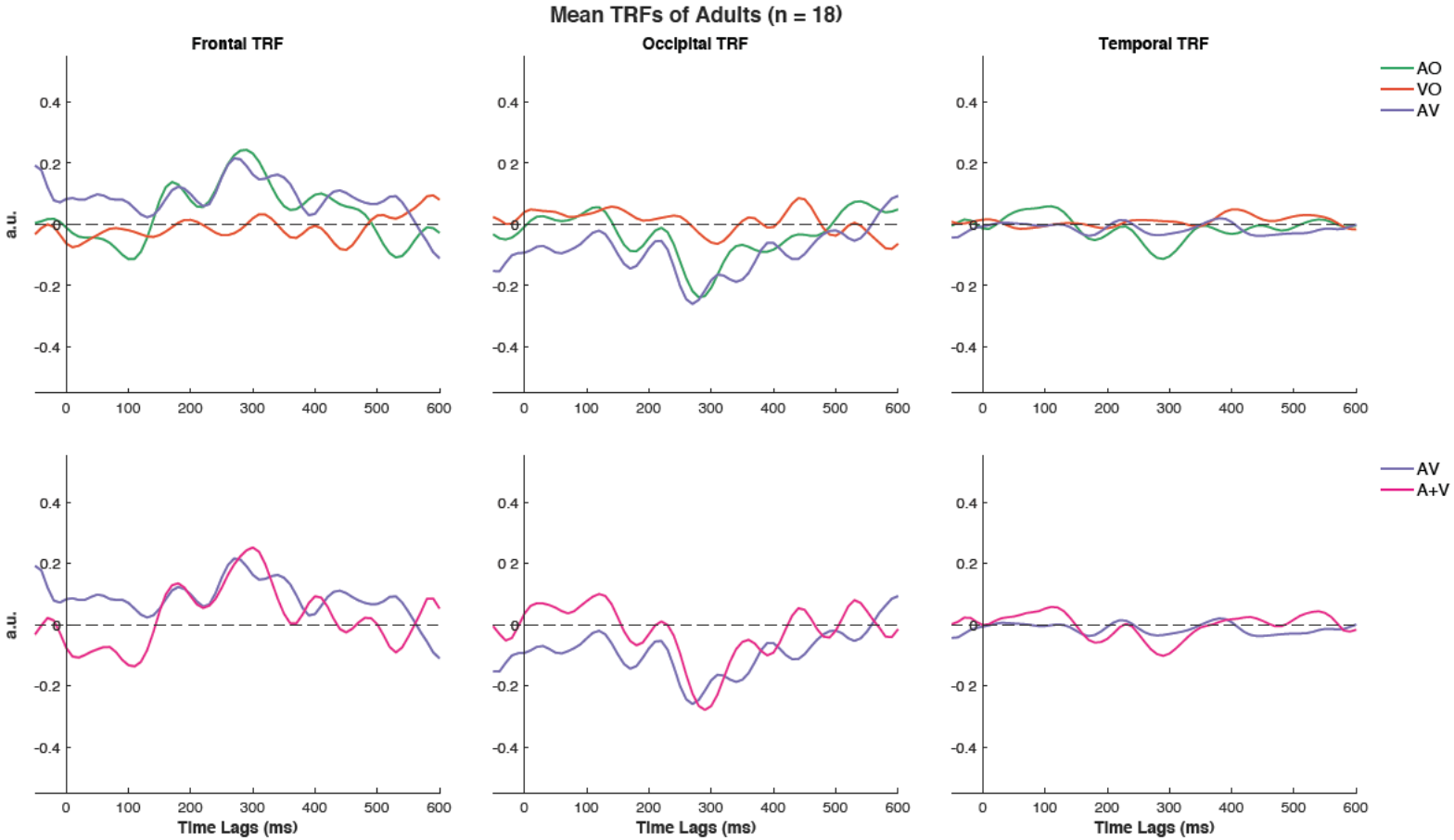
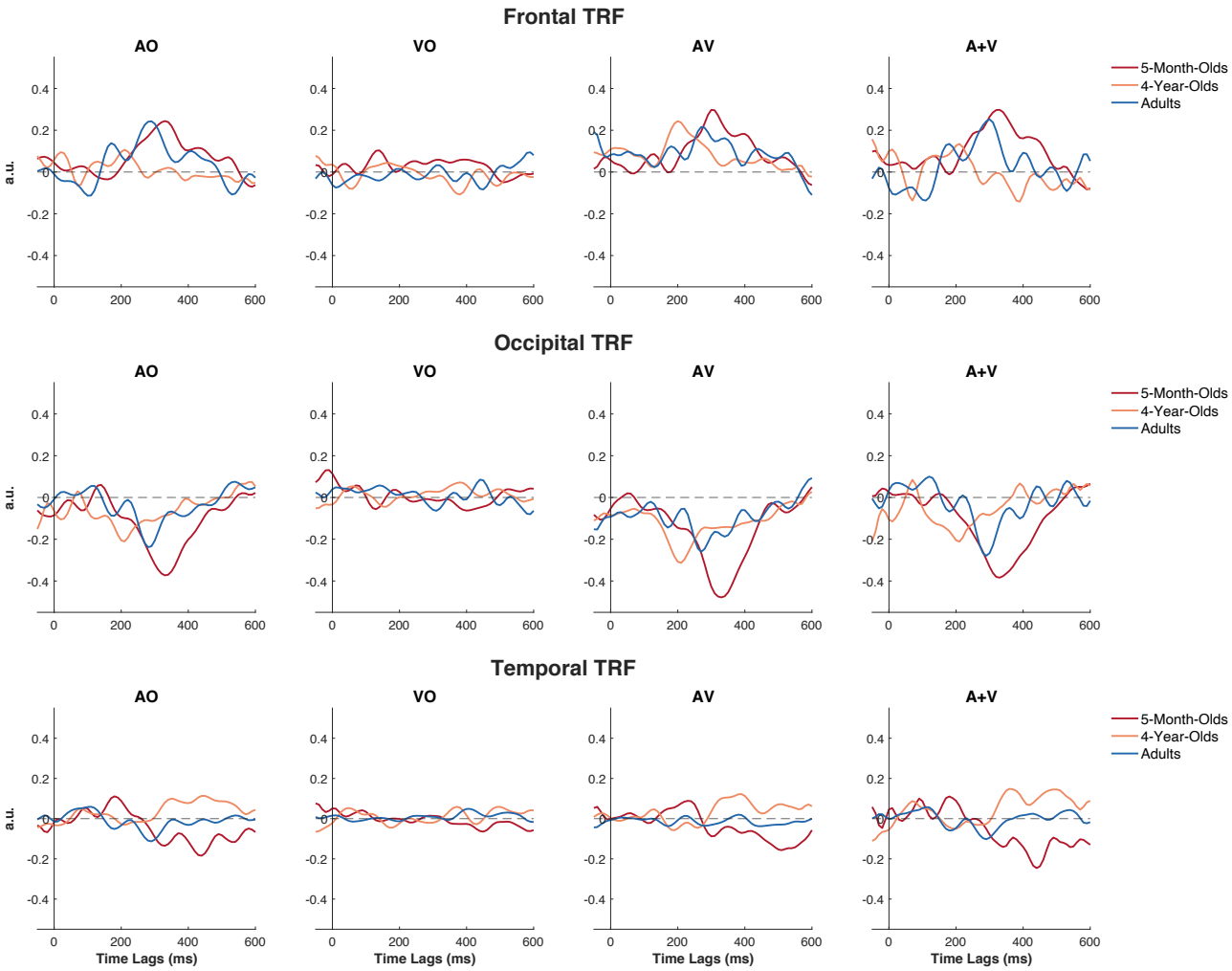


Figure 10

Temporal Response Functions for All Age Groups at the Three Scalp ROIs



3.3 Discussion

Study 1 examined the visual speech benefit 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 decoders (backward modelling approach) and temporal response functions (forward modelling approach). The backward modelling approach involves mapping data from all EEG channels simultaneously to decode stimulus features (broadband envelope) from the neural response, whereas the forward modelling approach involves mapping the speech envelope separately to each EEG electrode. Although the backward modelling approach may be a more sensitive measure of cortical tracking (Crosse, Di Liberto, Bednar, & Lalor, 2016), the forward modelling approach provides complementary spatiotemporal information because temporal response functions, unlike decoders, are neurophysiologically interpretable. For this study, stimulus reconstruction accuracy was first assessed to examine how well the cortical response in each condition was able to encode the speech envelope. Next, temporal response functions and their prediction accuracies were analysed in order to inspect the spatial and temporal profiles of the neural response to the stimuli. Visual speech benefit was inferred from the difference between neural responses to auditory-visual stimuli and the summation of neural responses to unimodal auditory-only and visual-only stimuli, i.e., [AV vs. (A+V)].

Results from the backward modelling approach indicate that cortical tracking of the speech envelope was weakest in the VO condition across all three age groups. Five-month-olds showed stronger cortical tracking in AV than AO, however there was no such difference was found for 4-year-olds or adults. Several interesting findings emerged from age comparisons. First, stimulus reconstruction accuracy was higher for 5-month-olds than 4-year-olds in all but the AO condition. Second, stimulus reconstruction accuracy was greater for 5-month-olds than adults in all conditions, suggesting that the cortical responses of 5-

month-olds encoded the speech envelope better. Third, stimulus reconstruction accuracies were similar for 4-year-olds and adults in every condition. Surprisingly, there was no indication of visual speech benefit—stimulus reconstruction accuracies were not significantly different between AV and (A+V) decoders at any of the three age groups. It must be considered that these age differences may be biophysical (i.e., skull thickness vary across ages) rather than neurocognitive, so greater stimulus reconstruction accuracies in 5-month-olds than 4-year-olds and adults may also stem from a stronger signal on their scalps. More work developing methods that would allow adjustments for biophysical differences between ages is necessary to disentangle these possibilities.

To understand the spatial and temporal dynamics of the stimulus-response mapping, temporal response functions were modelled and their prediction accuracies examined. Estimates of global field power illustrate clear components for AO, AV and (A+V) TRFs for all age groups. These components were generally larger, longer and later for 5-month-olds than 4-year-olds and adults. Topographic maps of TRF weights illustrate that most activity occurred during ~200-400ms at frontal/central, occipital and temporal scalp regions. Several age differences emerged with further examination of the TRFs at the three scalp ROIs. First, although all three age groups showed prominent frontal positivity and occipital negativity for AO, AV and (A+V) TRFs, these components appeared later and were slightly longer for 5-month-olds than for 4-year-olds and adults. Adult response functions were also more clearly defined compared to infant and child groups, suggesting that adult neural responses were more temporally modulated. Surprisingly, adult responses were generally later than those of 4-year-olds.

Age differences were also evident when prediction accuracies of the TRFs were examined. Specifically, 4-year-olds differed from 5-month-olds and adults in two ways. First, when prediction accuracies of AO, VO, and AV TRFs (averaged across all electrodes at the

three scalp ROIs) were compared, unlike 5-month-olds and adults, the prediction accuracy of AV TRFs was not significantly greater than the prediction accuracy of AO TRFs for 4-year-olds. Second, when 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 visual speech benefit for 5-month-olds—prediction accuracy of AV TRFs was significantly greater than the prediction accuracy of (A+V) TRFs—but not for 4-year-olds. The observed differences cannot be attributed to attention because 4-year-olds attended more on AV trials than on AO trials or VO trials (AV vs. AO: $t(18) = 6.10$, $p < .001$; AV vs. VO: $t(18) = 9.19$, $p < .001$). 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—but this was not the case.

Taken together, results from backward and forward modelling suggest that neural responses from 5-month-olds, 4-year-olds, and adults reliably track the speech envelope when presented with AO and AV speech. Importantly, age comparisons reveal that (1) there is effective cortical tracking of auditory-only and auditory-visual speech but not visual-only speech, and that (2) cortical tracking of auditory-visual speech becomes more localized to the temporal scalp regions. Across ages, visual speech alone was not sufficient to elicit reliable cortical tracking.

There were a number of unexpected findings. First, backward modelling showed that cortical tracking to AV speech did not differ from that of AO speech for adults. This was unexpected because previous behavioural (Moradi et al., 2013) and neurophysiological (Crosse, Di Liberto, & Lalor, 2016; Kaganovich & Schumaker, 2014; A. O'Sullivan et al., 2019) studies with adults have consistently found an AV>AO effect. It must be noted, however, that the difference between adults' AO and AV stimulus reconstruction accuracy approached significance ($p = .052$), raising the possibility that there may be a weak AV>AO

effect. Second, although the backward modelling approach did not indicate any effect of visual speech benefit, the forward modelling approach showed that cortical tracking of the speech envelope was stronger for AV TRFs than (A+V) TRFs at frontal, occipital, and temporal scalp locations for 5-month-olds and adults. This finding suggests that the visual speech benefit may be localised in these regions, and while the backward modelling approach provides a better stimulus-response fit to complex stimuli with low signal-to-noise ratio like speech (Crosse, Di Liberto, Bednar, & Lalor, 2016), the finding from backward modelling that greater stimulus reconstruction accuracy for AV than (A+V) decoder approached significance ($p = 0.06$) in adults along with results from forward modelling allow for a tentative conclusion that visual speech benefit is present in adults. It is also worth considering that the sample sizes (n = 18) for infants and adults may have resulted in the experiment being under-powered. It would be interesting in future studies to increase the sample size to investigate whether the subtle effects of AV vs. (A+V) are more evident. Notably, 4-year-olds, unlike the other two age groups, did not show any evidence of AV>AO effect or visual speech benefit. Moreover, these null results cannot be attributed to the differences in 4-year-olds' attention between conditions. It can be argued that attention to the screen may not be a precise measure of attention general, however, the absence of any derived benefit from the addition of visual speech information 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). As the IDS stimuli used in this study were constructed to cater to infants, it is possible that 4-year-olds processed the auditory speech quite effectively, such that the addition of visual speech information did not augment their cortical tracking.

Altogether, while only the results from the adult group consistently met the additive model criterion, the findings that prediction accuracy and stimulus reconstruction accuracy

were greater for AV speech than for AO speech, and that prediction accuracies of frontal, occipital and temporal scalp regions met the additive model criterion for 5-month-olds suggest that visual speech information enhances auditory speech processing to some extent. More investigations are necessary to probe the seemingly absent AV>AO effect observed in 4-year-olds.

To further explicate the present findings, two directions for future research are possible.

First, results from 4-year-olds call for more investigations to be conducted. While inconsistent findings of the 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 surprise, 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 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, despite the fact that the IDS stimuli may have been easy for them to process, the experiment was relatively lengthy (~25 minutes in total), and so 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 understand the speaker and this might have consequently affected their speech processing (Pichora-Fuller et al., 2016).

To address this issue, a viable modification of the current paradigm could be to use fewer but longer stimuli. For example, three 2-minute videos of a speaker reciting children stories can be presented once per condition (a total of 18 minutes)—as opposed to the thirty 8- to 15-second 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 mTRF approach especially since adult studies had larger and longer data sets (e.g., Crosse, Di Liberto, & Lalor (2016) used 15 x 60-s passages per condition). However, these concerns are allayed by a recent demonstration that the mTRF approach can be used effectively even with 7-month-olds' EEG data to a single 4-minute 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. (In the next section, 3.4, adults' cortical tracking of ADS is added and compared with their cortical tracking of IDS. This goes some way to addressing this issue, but to be complete, data on infants' cortical tracking would also be required.)

Results from this study provide evidence of a visual speech benefit in adults' cortical tracking of the speech envelope, in line with previous studies (e.g., Crosse et al., 2015). Future work is necessary to expound the small effect of visual speech benefit and the absence of any facilitation by visual speech cues in 4-year-old participants. Critically, this study demonstrates, for the first time, that the provision of visual speech cues increases the

accuracy with which 5-month-olds' neural oscillations synchronise with the speech envelope. In so doing, it lays the foundational brickwork for future investigations that seek to understand the neural mechanisms driving the visual speech benefit across development.

3.4 Addendum: Adults' Neural Responses to IDS vs. ADS

The exaggerated acoustic pitch (Kitamura et al., 2001) and prosody (Fernald & Mazzie, 1991) that are characteristic of infant-directed speech (IDS) are often paired with exaggerated facial expressions (Chong et al., 2003) and articulatory lip movements (Green et al., 2010). Compared to adult-directed speech (ADS), these exaggerated auditory and visual speech cues may come across as unnatural to adults. Further, a previous study found different event-related potentials (ERPs) in adults' responses to IDS and ADS indicating that their cortical processing of IDS differed from that of ADS (Peter et al., 2016). To rule out the possibility that the IDS stimuli used in the current study may have confounded the findings from adult data (reported previously in Section 3.2), adult participants were additionally presented with ADS stimuli.

3.4.1 Methods

3.4.1.1 Participants.

Sixteen of the 18 adult participants in the above study were included in this set of analyses. Data from 2 other participants were excluded because they did not meet the gaze criterion. The mean number of trials (per condition) included the analyses is 25.50, and the mean level of attention across conditions is 81.12%.

3.4.1.2 Stimuli.

Other than the type of speech and the word 'baby' removed from two speech passages (1 and 27; Appendix C) for appropriateness, everything else was the same as described in Section 3.1.2. Auditory-visual recordings were made of the same female native speaker of Australian English reciting the same 30 short speech passages, this time in adult-directed speech. Unimodal auditory and visual recordings were extracted separately for AO and VO conditions. The speech passages varied in durations from 7.20s to 14.20s (mean = 9.96s, SD = 1.54s). As before, the 30 speech passages were presented in three blocks with presentation

order randomised across modality and in such a way that the same sentence did not appear in two modalities on consecutive trials. Attention-getter stimuli were the same pictures of '*Minions*' that appeared in a random order after either two or three trials, with their frequency randomly determined. A 3-s cartoon animation was played to mark the end of each block and to re-engage participants.

3.4.1.3 Procedure.

The procedure is identical to that described in Section 3.1.3.

3.4.2 EEG Pre-Processing and Data Analysis

The same pre-processing pipeline and data analysis methods as described in Section 3.1.4.1 were used here.

3.4.3 Results (IDS vs. ADS)

The same data analysis procedure as described for IDS data (Section 3.2) was applied here. First, decoders, or backward models, were computed from neural responses to ADS. Stimulus reconstruction accuracies of these decoders were compared against (1) decoders constructed by mapping neural responses to shuffled ADS envelopes, and (2) decoders computed from mapping neural responses to IDS. Next, TRFs, or forward models, were modelled from neural responses to ADS. Only prediction accuracies from TRFs generated at the scalp ROIs used for IDS analyses (frontal, occipital and temporal locations) were used in the analyses here.

3.4.3.1 Decoders.

A Condition (AO vs. VO vs. AV vs. A+V) x Envelope Type (Original vs. Shuffled) repeated-measures ANOVA for ADS speech to examine if stimulus reconstruction accuracy was different between conditions. The two main effects (Condition and Envelope Type) and the Condition x Envelope Type interaction were significant (Condition: $F(3, 45) = 7.98$, $p < .001$, Envelope Type: $F(1, 15) = 23.94$, $p < .001$, $\eta_p^2 = .61$, Condition x Envelope Type: $F(3,$

45) = 7.19, $p < .001$, $\eta_p^2 = .32$). Separate one-way repeated-measures ANOVA performed for each Envelope Type indicated that significant differences in stimulus reconstruction accuracy between conditions were present only for the original speech envelopes (original speech envelopes: $F(3, 45) = 11.12$, $p < .001$, $\eta_p^2 = .43$; shuffled speech envelopes: $F(3, 45) = 1.06$, $p = .37$, $\eta_p^2 = .07$). Planned paired-sample *t*-tests with Bonferroni correction revealed that reconstruction accuracy was greater for AO than for VO decoder ($t(15) = 5.18$, $p < .001$, Hedge's $g = 1.40$), and for AV than for VO decoder ($t(15) = 4.72$, $p < .001$, Hedge's $g = 1.28$). The difference between AO and AV decoders and the difference between AV and (A+V) decoders were not significant, (AO vs. AV: $t(15) = -0.47$, $p = .64$, Hedge's $g = -0.13$; AV vs. (A+V): $t(15) = 0.69$, $p = .50$, Hedge's $g = 0.12$).

To investigate the differences between stimulus reconstruction accuracy of decoders mapping neural responses to IDS speech and decoders mapping neural responses to ADS speech, a Speech Type (IDS vs. ADS) x Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted. Only the main effect of Condition was significant, $F(1, 15) = 14.40$, $p < .001$, $\eta_p^2 = .49$. The main effect of Speech Type and the Speech Type x Condition interaction was not significant (Speech Type: $F(1, 15) = 4.43$, $p = 0.053$, $\eta_p^2 = .23$, Speech Type x Condition: $F(1, 15) = 1.71$, $p = 0.18$, $\eta_p^2 = .10$).

Although the main effect of Speech Type and Speech Type x Condition interaction were not significant, planned paired-sample *t*-tests were conducted to further examine if there was any difference between stimulus reconstruction accuracy of IDS and ADS decoders in the three conditions. Paired samples *t*-tests with Bonferroni correction showed that reconstruction accuracy of AO, VO and AV decoders did not differ between IDS and ADS speech (AO: $t(15) = -.244$, $p = 0.028$, Hedge's $g = -0.63$; VO: $t(15) = 0.23$, $p = 0.83$, Hedge's $g = 0.08$; AV: $t(15) = -1.13$, $p = 0.28$, Hedge's $g = -0.30$), suggesting that there was no difference in adult envelope tracking of IDS and ADS speech. All means and standard

deviations are reported in Table 3. Figure 11 depicts individual stimulus reconstruction accuracy values for each condition across speech types.

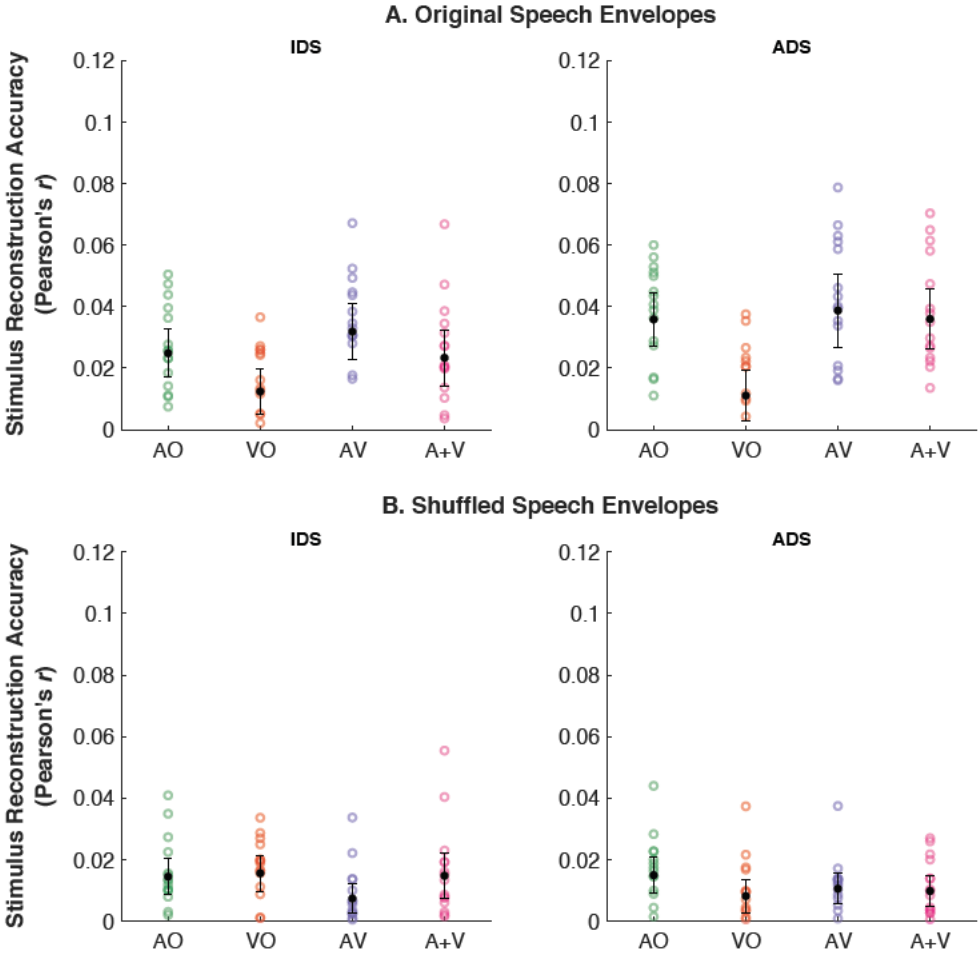
Table 3

Means (and Standard Deviations) of Stimulus Reconstruction Accuracies for IDS and ADS

	AO	**VO**	**AV**	**A+V**
Original Speech Envelopes				
IDS	.025 (.016)	.012 (.015)	.032 (.019)	.023 (.019)
ADS	.036 (.018)	.011 (.017)	.039 (.025)	.036 (.020)
Shuffled Speech Envelopes				
IDS	.015 (.012)	.016 (.012)	.007 (.010)	.015 (.015)
ADS	.015 (.012)	.008 (.011)	.011 (.010)	.010 (.010)

Figure 11

Stimulus Reconstruction Accuracy for IDS and ADS



Individual stimulus reconstruction accuracy values using Pearson's correlation coefficient for adults' (n = 16) responses to IDS and ADS when decoders were trained and tested on (a) the original speech envelopes, and (b) the shuffled speech envelopes. Mean accuracies with 95% confidence intervals are shown in black.

3.4.3.2 Temporal Response Functions.

Similar to IDS, an estimate of GFP was calculated. The temporal profile of GFP for ADS depicted clear components for AO, VO, AV and (A+V) TRFs (Figure 12). This is in contrast to the GFP for IDS, where no clear component was observed for VO TRF. For visualization purposes, Figure 13 depicts the topographies of TRF weights for ADS, Figure 14 illustrates the mean frontal, occipital, and temporal TRFs for ADS, and Figure 15 shows the temporal response functions of both speech types at the three scalp ROIs.

To make direct comparisons with results from adults' cortical responses to IDS, the same frontal, occipital and temporal scalp ROIs were used here. One-tailed one-sample *t*-tests were used to assess the *presence* of cortical tracking at these regions by testing the respective mean prediction accuracies against zero. To examine whether the strength of cortical tracking differed as a function of *condition*, a one-way repeated-measures ANOVA was conducted with prediction accuracy as the dependent variable and model type as the independent variable. To examine whether the strength of cortical tracking differed as a function of *speech type*, a Speech Type (IDS vs. ADS) x Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted on prediction accuracy values.

3.4.3.2.1 Evidence of cortical tracking.

One-tailed one-sample *t*-tests revealed that only the mean prediction accuracy of AO and AV TRFs at the scalp ROIs were significantly greater than zero (AO: $t(15) = 4.82$, $p = .008$, Hedge's $g = 1.14$; AV: $t(15) = 2.17$, $p = .046$, Hedge's $g = 0.52$). Prediction accuracy of VO and (A+V) TRFs were not significantly greater than zero (VO: $t(15) = -1.86$, $p = .08$, Hedge's $g = -0.44$; A+V: $t(15) = 1.81$, $p = .09$, Hedge's $g = 0.43$).

3.4.3.2.2 Differences in strength of cortical tracking.

The repeated-measures one-way ANOVA was significant, $F(3, 45) = 2.82$, $p = .049$, $\eta_p^2 = .16$, indicating that the differences between conditions were significant. To further

inspect these differences, paired-sample *t*-tests were conducted with Bonferroni correction. These *t*-tests revealed that the prediction accuracy of AO and AV TRFs were significantly greater than VO TRFs (AO vs. VO: $t(15) = 7.47$, $p < .001$, Hedge's $g = 1.74$; AV vs. VO: $t(15) = 3.12$, $p = .007$, Hedge's $g = 0.91$), but not significantly different from each other (AO vs. AV: $t(15) = 0.79$, $p = .44$, Hedge's $g = 0.21$). The prediction accuracy of AV TRFs were also not significantly different from (A+V) TRFs, $t(15) = 1.19$, $p = .25$, Hedge's $g = 0.14$, suggesting that visual speech benefit was not present.

To examine the differences between speech types, a Speech Type (IDS vs. ADS) x Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted. The main effect of Condition ($F(2, 30) = 25.47$, $p < .001$, $\eta_p^2 = .63$), and the Speech Type x Condition interaction were significant ($F(2, 30) = 4.47$, $p = .02$, $\eta_p^2 = .23$), but not the main effect of Speech Type ($F(2, 30) = 0.71$, $p = .41$, $\eta_p^2 = .05$). Planned post-hoc comparisons (IDS vs. ADS) conducted for each condition via pairwise *t*-tests with Bonferroni correction indicated that prediction accuracy did not differ as a function of speech type for any condition (AO: $t(15) = -1.09$, $p = .29$, Hedge's $g = -0.35$; VO: $t(15) = 0.87$, $p = .40$, Hedge's $g = -0.29$; AV: $t(15) = 1.93$, $p = .07$, Hedge's $g = 0.52$). Means and standard deviations of prediction accuracy for each Speech Type and Condition are reported in Table 4.

Figure 12

Global Field Power Measured at Each Time Lag for IDS and ADS

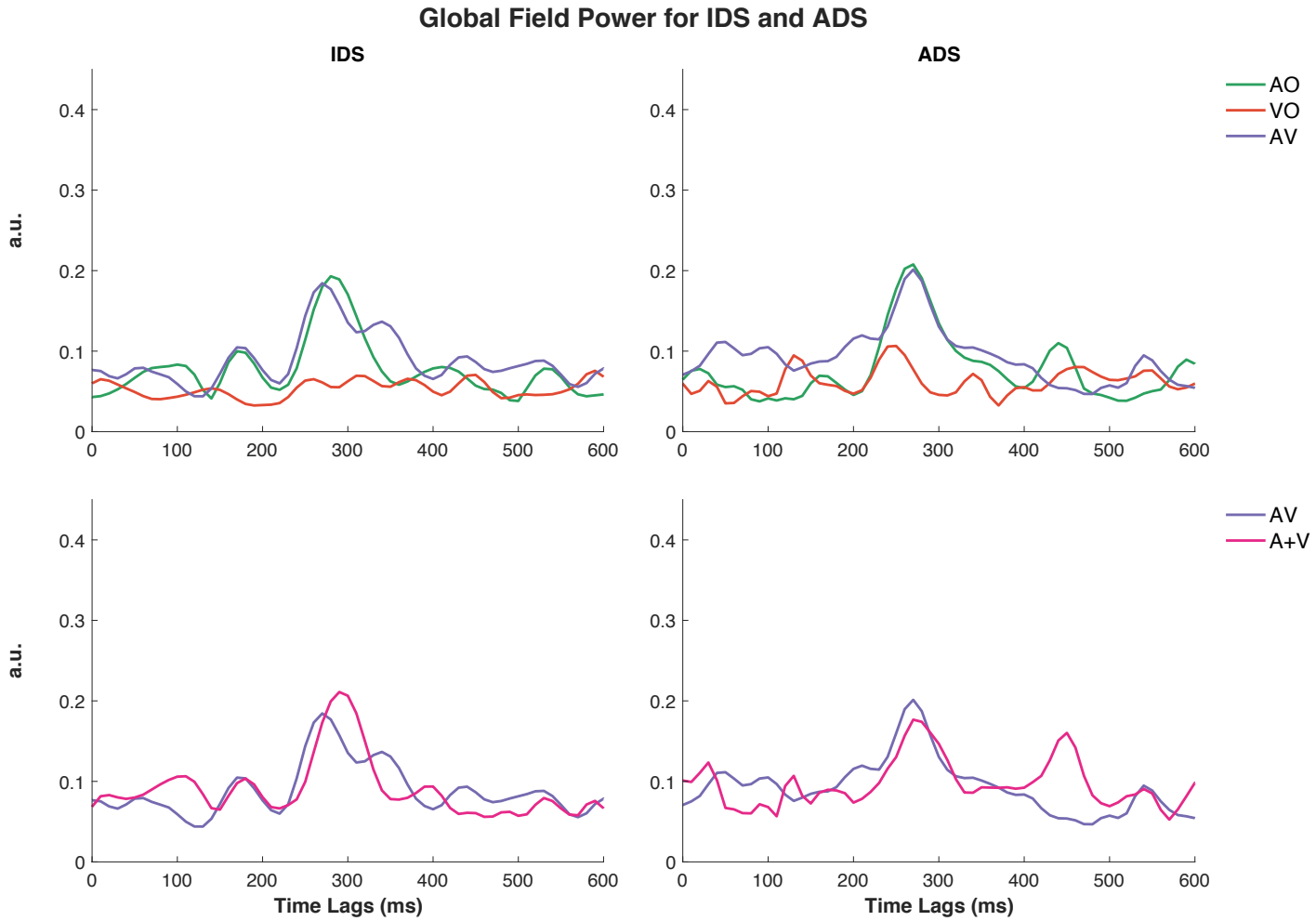
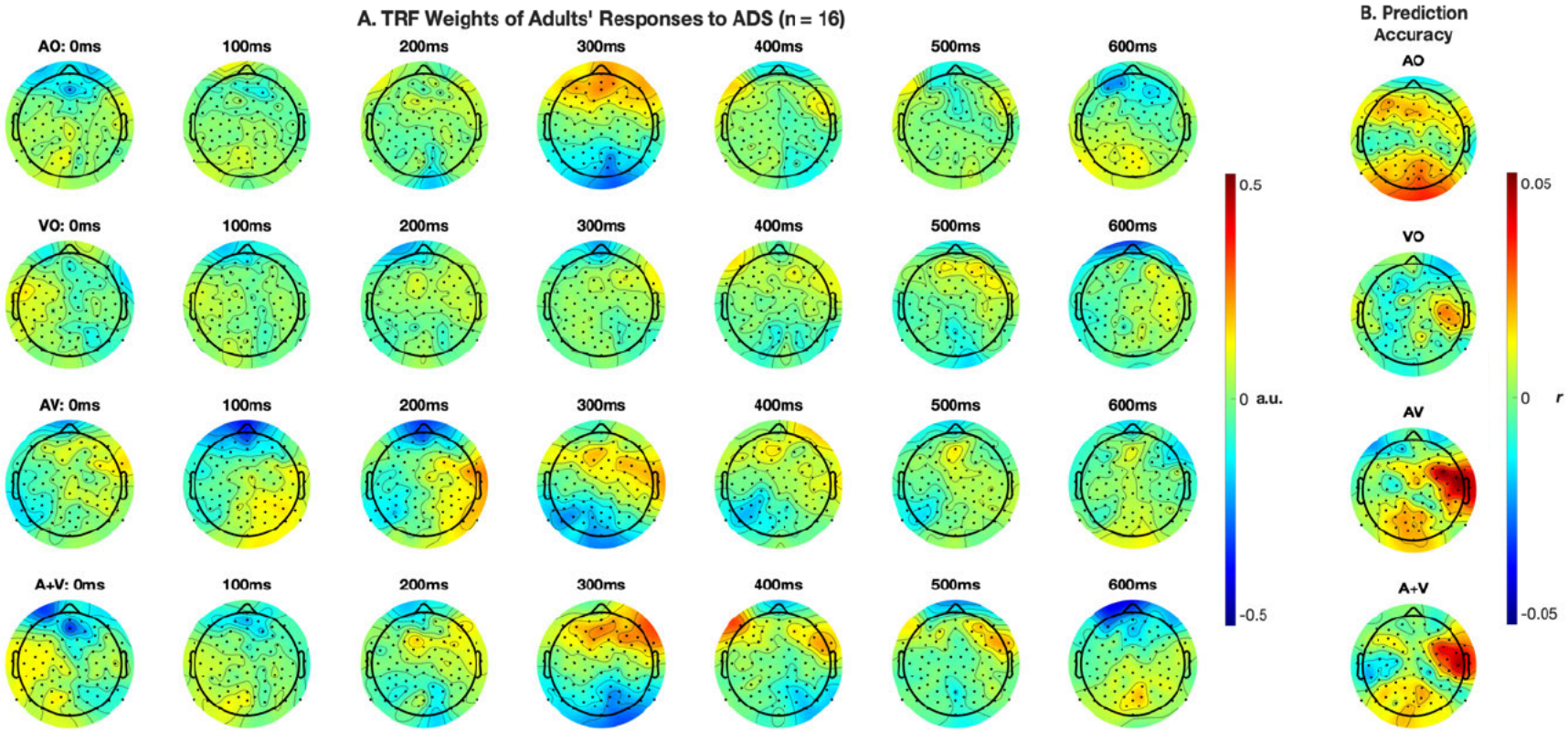


Figure 13

Topographies of TRF Weights for ADS



Topographies illustrate (a) TRF weights at time lags of 0-600ms with 100ms intervals, and (b) prediction accuracy values (Pearson's *r*) of the TRFs for each condition.

Table 4

Mean Prediction Accuracies (and Standard Deviations), Quantified by Pearson's r, of TRFs from Frontal, Temporal and Occipital Scalp ROIs for each Condition and Speech Type (IDS and ADS)

	AO	**VO**	**AV**	**A+V**
IDS	.010 (.012)	-.0004 (.011)	.020 (.014)	.005 (.011)
ADS	.014 (.012)	-.003 (.007)	.011(.019)	.008 (.017)

Figure 14

Temporal Response Functions for Adults' Neural Responses to ADS at Frontal, Occipital, and Temporal Scalp Locations

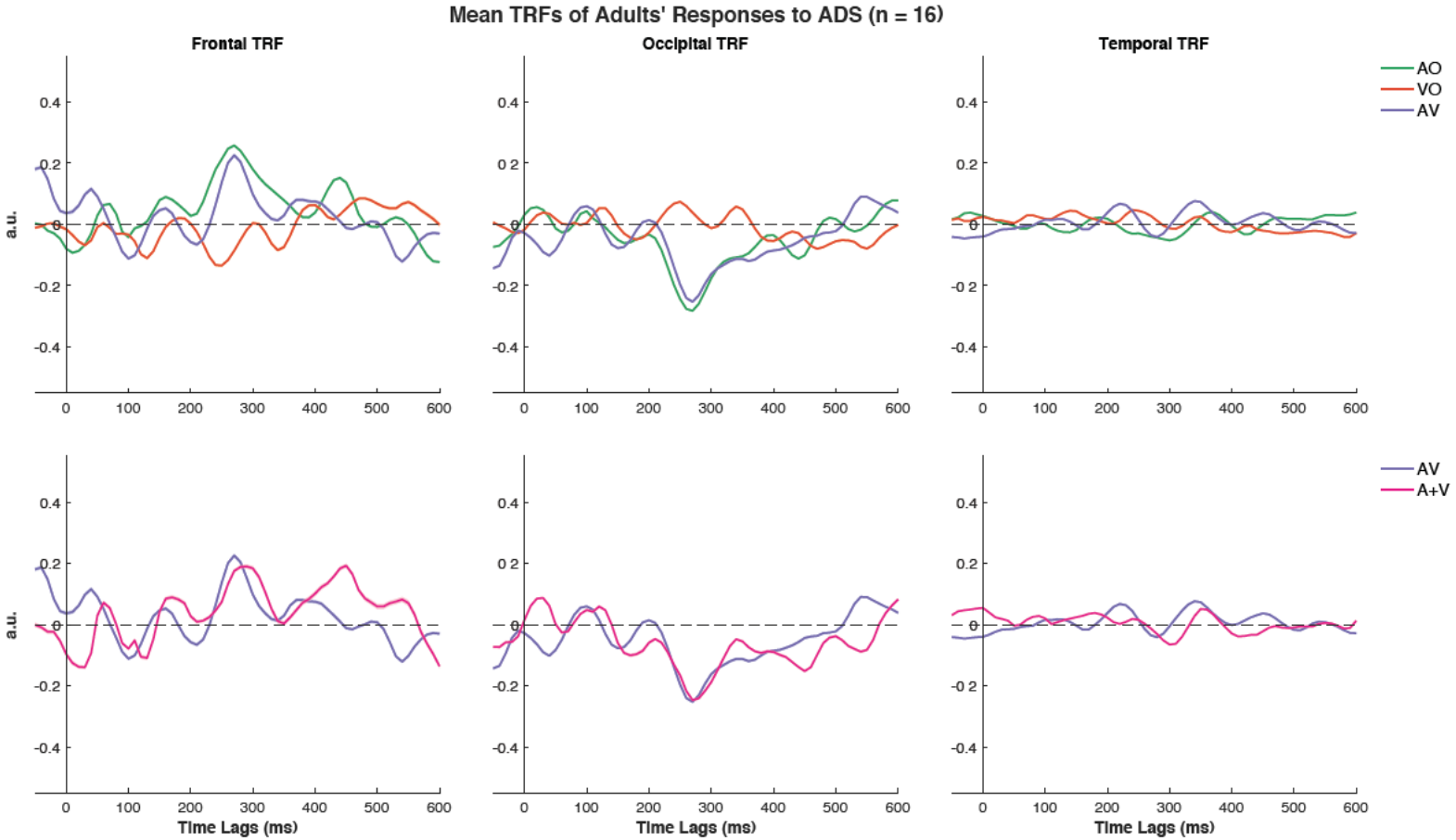
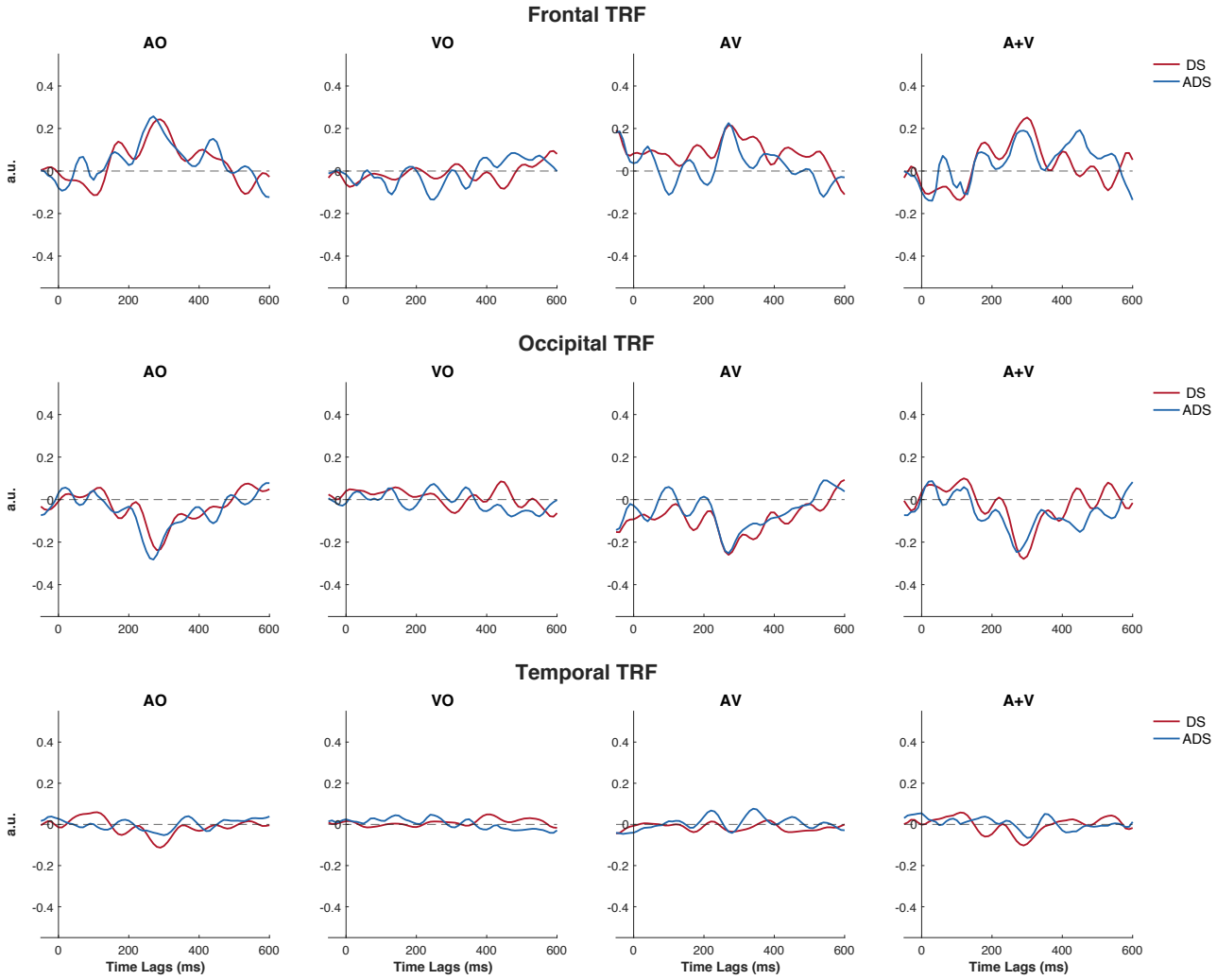


Figure 15

Temporal Response Functions for Adults' Neural Responses to IDS and ADS at Frontal, Occipital and Temporal Scalp Locations



3.4.4 Discussion

To address the possibility that adult envelope tracking differed as a function of speech type, adult neural responses to ADS were recorded and analysed. Differences emerged when the cortical tracking accuracy was compared between conditions. Specifically, results from the backward modelling approach indicated that the strength of adults' cortical tracking did not differ between unisensory modalities when IDS was presented, whereas cortical tracking was more accurate for AO than for VO speech when ADS was presented. Forward modelling analyses which focused on the three scalp ROIs (frontal, temporal, and occipital) showed that cortical tracking accuracy was higher for AV than for AO TRFs when IDS was presented, but this difference was not observed when ADS was presented. Visual speech benefit, as quantified by the additive model criterion, was present for IDS but not for ADS.

Two notable implications can be gleaned from these results. First, the provision of the speaker's talking face was more beneficial for adults' cortical tracking of IDS than ADS. This suggests that adults had more difficulty comprehending IDS compared to ADS—while the auditory ADS signal was alone was sufficient for adults' speech processing, this was not the case for IDS. The visual speech benefit observed only for IDS corroborates previous behavioural studies with adults showing increased reliance on visual speech cues (e.g., Birulés et al., 2020) and a greater visual speech benefit (e.g., Blackburn et al., 2019) in challenging listening conditions. Second, cortical tracking accuracy did not differ between IDS and ADS for any condition, indicating that the accuracy with which adults' cortical oscillations track the IDS envelope (as reported in Sections 3.2.1.3 and 3.2.2) was not atypical even though IDS may be unnatural to adults.

3.5 Summary

To investigate the visual speech benefit at the neural level, 5-month-olds, 4-year-olds, and adults were presented with short speech passages in auditory, visual, and auditory-visual

modalities as EEG and gaze data were simultaneously recorded. The visual speech benefit was evident for infants and adults but not for 4-year-olds. A secondary examination of adults' neural responses to IDS and ADS was conducted to rule out the potential confounding role that speech type may play. This investigation proved informative: in addition to demonstrating that adults' cortical tracking accuracy was not significantly different for IDS and ADS, it also revealed that, unlike IDS, there was no evidence of any visual speech benefit in adults' cortical tracking of ADS—a finding that supplements and corroborates behavioural studies that have shown an increase in adults' reliance on visual speech cues in difficult listening situations. Being the first study to examine the visual speech benefit in infants' and children's cortical tracking of the speech envelope, it brings to stark relief that there are more questions than answers. By doing so, this study sets the stage for future forays into understanding the neural mechanisms responsible for the visual speech benefit.

CHAPTER 4

Study 2: Infants' Segmentation of Continuous Auditory-Visual Speech

In Study 1, a visual speech benefit was found for 5-month-olds' cortical tracking of the speech envelope, a result that corroborates the few behavioural studies showing a visual speech benefit (see 2.2.3; Hollich et al., 2005; Teinonen et al., 2008). Studies investigating the visual speech benefit in infants are few in number. In this chapter, the visual speech benefit is examined specifically in one of the earliest tasks that infants must master for successful language acquisition; word segmentation—accurate identification of word boundaries and segmentation of utterances into words.

Word segmentation is by no means an easy feat—word boundaries are not systematically marked by acoustic cues (Aslin et al., 1996; Cole et al., 1980), no two utterances are exactly the same and speaker characteristics are highly variable. Despite the complex nature of speech, as shown in Chapter 2 there is substantial evidence that, regardless of linguistic background, infants segment continuous speech within the first year of life (e.g., Butler & Frota, 2018; Jusczyk & Aslin, 1995; Marquis & Shi, 2008) and this ability is further shaped by their native language, becoming increasingly robust with age (e.g., Schmale et al., 2010; Thiessen & Saffran, 2004).

Studies of infant word segmentation have largely focused on auditory-only speech even though speech perception is now known to be a multimodal process. Segmentation studies using artificial languages have shown that adults are better able to identify words within an artificial language stream when the stream is paired with a congruent dynamic video of a speaker's talking face than when it is presented only in the auditory modality (Lusk & Mitchel, 2016; Mitchel & Weiss, 2010; 2014). This raises the possibility that visual speech information may aid infants' word segmentation, especially since behavioural studies

have found that infants benefit from visual speech information in other speech perception tasks (e.g., Teinonen et al., 2008). There is, to my knowledge, only one study that has examined infants' auditory-visual word segmentation. In that study, Hollich et al. (2005) familiarised infants with passages presented in auditory-visual mode by a female speaker. The auditory component of these passages was blended with distractor auditory-only passages recited by a male speaker. In subsequent test trials 7.5-month-olds listened longer to auditory tokens of familiar than unfamiliar words spoken by the female. Thus, they were able to separate the two voices in familiarisation and attend to the female voice, presumably because the visual information was temporally synchronous with the female but not the male voice. This explanation was strengthened by further experiments that showed that infants did not segment target words successfully when the female's auditory passages in familiarisation were paired with a static image of her face or with an unsynchronised video display of her speaking. They were successful, however, when the visual display was an oscilloscope pattern synchronised with the female's voice. These results show that visual information from a speaker's talking face enhances infants' segmentation of continuous speech in the presence of background noise and suggests that this enhancement is related to the temporal concordance of the auditory and visual information. No subsequent study has investigated whether the same augmentation can be found even in optimal listening conditions where the auditory signal is clear (without any distractor background speech).

To derive benefits from an interlocutor's talking face, infants must first *attend* to the talking face. Research on gaze behaviour to the eye and mouth regions of a speaker's talking face has shown that infants treat the mouth as an important source of linguistic information: infants already seek linguistic information from the mouth within the first half of the year (Tenenbaum et al., 2013), will attend more to the speaker's mouth than the eye region at 8 months when the speaker talks in a native or a non-native language (Lewkowicz & Hansen-

Tift, 2012), and continue to do so at 12 months but only when the speaker talks in a non-native language (Pons et al., 2019). Together, these findings demonstrate that infants can efficiently deploy their attention to relevant visual speech cues.

What is not clear from infant word segmentation research is whether there is a direct link between infant gaze behaviour, specifically fixation to the mouth, and their word segmentation performance. It is possible that individual differences in directing gaze to the speaker's mouth may modulate infants' word segmentation performance. As the temporal pattern of mouth movements is highly correlated with the acoustic timescale of syllables, researchers have proposed that mouth movements provide information regarding the onset and offset of syllables (Chandrasekaran et al., 2009) and that infants can make use of the alignment between auditory and visual components of speech such as mouth movements to segment speech (Kitamura et al., 2014). Nevertheless, the relationship between infant gaze behaviour and word segmentation has not yet been studied directly.

Study 2 addresses the modulating effects of visual speech information on infant word segmentation and the possible gaze pattern mechanisms that might underlie this instance of the visual speech benefit. There are two aims: (1) to examine whether visual speech information enhances infants' segmentation performance even in the absence of background distractor speech here (in contrast to the stimuli used in Hollich et al., 2005); and (2) to assess whether individuals' differential gaze to the mouth and eye regions modulates segmentation performance. As in previous segmentation studies that employed the passage-to-word paradigm (e.g., Hollich et al., 2005; Jusczyk & Aslin, 1995; see also Section 2.4.2), the 7.5-month-old infants in Study 2 here were first familiarised with passages containing target words and then tested with isolated tokens of target and non-target words. One group of infants was presented with familiarisation and test stimuli in the auditory-visual modality while a second group of infants was presented with stimuli in the auditory-only modality.

Unlike previous segmentation studies that have traditionally used the head-turn preference paradigm, this study used a familiarisation-test procedure with a single central screen (as in Thiessen, 2010) to accommodate the use of an eye-tracker to record infants' gaze patterns.

The specific hypotheses are:

1. If visual speech information augments word segmentation, then infants in the auditory-visual condition will show better segmentation performance than infants in the auditory-only condition, and
2. If the speaker's mouth is the main source of visual speech benefit, then attention to the mouth region and/or attention to the mouth compared to the eye region will be positively correlated with segmentation performance by infants in the auditory-visual condition (but not in the auditory-only condition).

4.1 Methods

4.1.1 Participants

Thirty-seven 7.5-month-olds monolingual Australian-English learners (20 females, mean age = 7.21 months, range = 7.03-7.90 months) were recruited. Eighteen infants participated in the auditory-only (AO) condition (9 females, mean age = 7.38 months, range = 7.03-7.87 months), and 19 participated in the auditory-visual (AV) condition (11 females, mean age = 7.19 months, range = 7.07-7.90 months). All infants were born full-term, with no vision or hearing deficits, and were not at risk for any language or cognitive delay and had no history of ear infections. Data from eight additional infants were excluded because they were either fussy and failed to complete the experiment (n = 6) or had less than 40% weighted gaze samples (n = 2). 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.

4.1.2 Stimuli

A female native Australian English speaker was recorded producing, in infant-directed speech, the four different 6-sentence passages used by (Jusczyk & Aslin, 1995) (Appendix D). Passages centred around the target words 'cup', 'dog', 'bike', and 'feet'. The recordings were auditory-visual of the speaker's head, face and neck. The average duration of the passages was 24.62s. Additionally, for the single word test stimuli the speaker spoke each of the four target words 8 times in succession incorporating some degree of variation in intonation. In the auditory-visual (AV) condition, these video recordings, consisting of a speaker's talking face and the audio recordings were used. In the auditory-only (AO) condition auditory recordings were extracted from the video recordings and paired with a still image of the speaker's smiling face.

4.1.3 Apparatus

Video recordings were presented via a 17-inch DELL LCD monitor and auditory recordings were played via two loudspeakers (Edirol MA-15 Digital Stereo Micro Monitors) placed at the left and right side of the monitor. A Tobii X120 eye tracker was placed below the bottom of the screen to record infants' gaze patterns throughout the session. The eye movements of each infant were calibrated using a 5-point calibration routine before the session began.

4.1.4 Procedure

A familiarisation-then-test design was employed using a single central screen (Thiessen, 2010). Infants sat on their parent's lap approximately 60cm from the screen. An experimenter was stationed in the adjacent control room throughout the experiment. The experiment consisted of two familiarisation and two test phases. In each phase, stimuli were played until completion. A schematic representation of the procedure is shown in Figure 16a.

In Familiarisation Phase 1, infants were presented with two repetitions of two passages, a total of four trials. Half of the infants were familiarised with two repetitions of the passages containing 'cup' and 'dog' while the other half were familiarised with two repetitions of the passages containing 'bike' and 'feet'. The two passages were presented on alternate trials, e.g., 'cup' repetition 1, 'dog' repetition 1, 'cup' repetition 2, 'dog' repetition 2, with order of passages counterbalanced between infants. The mean duration of this familiarisation phase was 98.90s.

In Test Phase 1, all four words (*cup*, *dog*, *bike*, and *feet*) were presented to the infants either as targets or non-targets. Target words are the words that appeared in the passages, i.e., for infants who heard passages that contained 'cup' or 'dog', 'cup' and 'dog' are target words while 'bike' and 'feet' are non-targets. In a given test trial, infants heard eight different tokens of a single word (e.g., eight repetitions of 'cup'). The test trials alternated between target and non-target words, with order of target and non-target words counterbalanced between infants (i.e., the first test trial was a target word for some infants and for other infants a non-target). The test phase consisted of two blocks of four trials—each block contained each of the two target and the two non-target words, resulting in a total of 8 trials. Each test block of target and non-target trials was on average 64.22s (therefore, the duration of Test Phase 1 is approximately 128.44s).

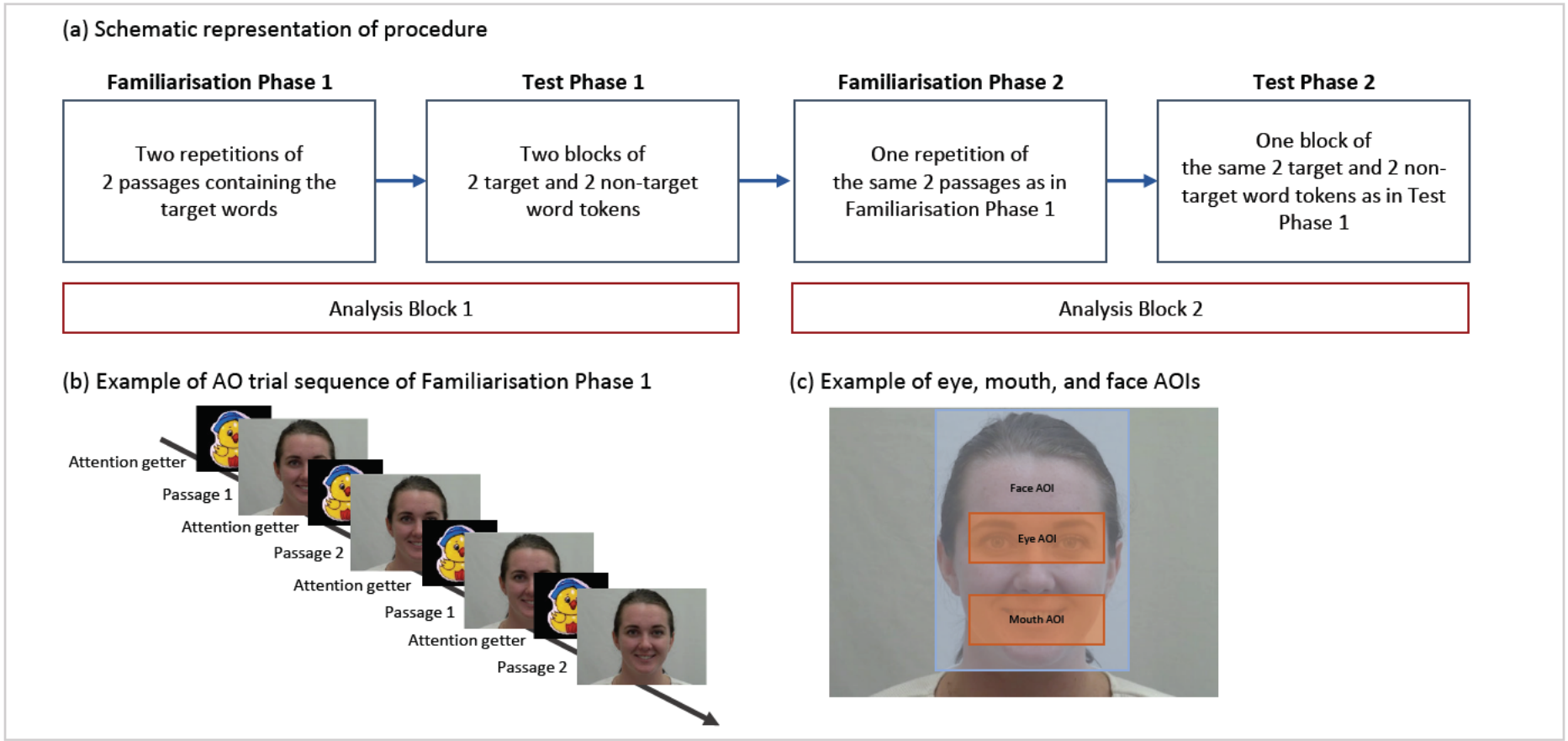
In Familiarisation Phase 2, the two familiarisation passages were presented once more (mean duration of this phase was 49.45s). Following this, in Test Phase 2, a single block of test trials was presented (see Figure 16a). An attention-getter animation was played in between trials and phases in order to recapture infants' attention to the screen (see Figure 16b).

Phase 2 (Familiarisation and Test) was added to the paradigm in order to ensure that infants had sufficient exposure to the familiarisation passages. In typical word segmentation

studies, a version of the head turn preference paradigm is usually employed in which familiarisation stimuli continue to play until the infant reaches a certain listening criterion (e.g., Jusczyk & Aslin, 1995; Hollich et al., 2005). That method is used to ensure that infants actively listen for a certain duration—for example, the familiarisation phase in the Jusczyk and Aslin study (1995) continued until the infant accumulated at least 45s of listening time per familiarisation passage. Implementing a listening criterion during familiarisation was not possible in this study because, unlike previous segmentation studies, looking behaviour is of interest in this study. To investigate gaze behaviour, it was important that stimulus presentation is kept constant for all infants, and that the overall duration of stimuli in each phase allowed for an adequate amount of gaze data to be captured. As such, Familiarisation Phase 2 and Test Phase 2 were added in this study to ensure that infants have a sufficient amount of exposure to the familiarisation passages. A single block of test trials was used in the second phase in order to minimise the overall duration of the test session.

Figure 16

(a) Schematic representation of the procedure in Study 2, (b) an example of a trial sequence (Familiarisation Phase 1) illustrating an attention getter preceding each trial, and (c) an example of the dynamic eye, mouth and face AOIs that were defined for each trial (and frame for AV stimuli)



4.1.4.1 Auditory-Only Condition.

Trials were initiated once infants attended to the attention getter. During familiarisation, a static image of the speaker's face appeared on the screen while the auditory recordings of the passages were played through to completion regardless of infant gaze behaviour. The static image faded from the screen at the end of each auditory recording.

Each test phase began immediately after each familiarisation phase. Once the infant attended to the attention getter, the experimenter initiated the test stimuli. The auditory recording of the word tokens paired with a static image of the speaker's face played until completion. The trial ended with the static face fading from the screen. Looking times to the screen during test trials were recorded as an index of infants' relative preference for the target and non-target words.

4.1.4.2 Auditory-Visual Condition.

The procedure for the auditory-visual condition was identical to the AO condition except that familiarisation and test stimuli were presented in auditory-visual modality—dynamic videos of the speaker reciting the passages (in familiarisation) and word tokens (in test) were presented.

4.1.5 Eye-Tracking Analyses

Dynamic areas of interest (AOIs) for the speaker's eye, mouth and face regions were defined (Figure 16c). The AOIs for the eye and mouth regions are of equal sizes (343 by 153 pixels). Looks to these three defined regions and looks to the screen in general were collected and recorded in Tobii Studio v 3.4.5.

Raw looking times were extracted using *dplyr* (Wickham, Francois, Henry, & Muller, 2020) and *tidyr* (Wickham & Henry, 2020) packages in custom R scripts (R Core Team, 2020). Proportions of total looking times (PTLs) to these regions were then calculated from the raw looking times for each familiarisation and test trial. The PTLs of interest are:

(1) mouth vs total looks $[\frac{fixation\ duration\ to\ mouth}{total\ fixation\ duration\ to\ screen}]$,

(2) eyes vs total looks $[\frac{fixation\ duration\ to\ eyes}{total\ fixation\ duration\ to\ screen}]$, and

(3) face vs total looks $[\frac{fixation\ duration\ to\ face}{total\ fixation\ duration\ to\ screen}]$,

Additionally, attention on each trial was derived as $[\frac{total\ fixation\ duration\ to\ screen}{trial\ duration}]$.

All statistical analyses on PTLs were conducted in R (R Core Team, 2020).

A drawback of using PTL is that it does not provide us with information on how infants' looking behavior may vary across the trial time window. It may be the case that infants shift their gaze from one AOI to another, but such shifts are not reflected in PTL calculations. Therefore, additional time-course analyses were performed with the *eyetrackingR* (Dink & Ferguson, 2018) package to examine whether differences in looking behavior between the two groups of infants emerge across time.

Familiarisation Phase 1 and Test Phase 1 were analysed as Block 1, and Familiarisation Phase 2 and Test Phase 2 were analysed as Block 2.

4.2 Results

Data were analysed in four parts. First, preliminary analyses were conducted to examine whether attention during familiarisation differed between the two groups of infants. Second, word segmentation was examined by comparing attention to target and non-target test trials. Third, regression analyses were conducted to investigate whether looking behaviour during familiarisation predicts segmentation performance in test. Finally, time course analyses were conducted to explore infants' word segmentation as a function of their looking behaviour to face, eyes, and mouth over time.

4.2.1 Attention During Familiarisation

Mean and standard deviations (SD) of attention are reported in Table 5 for both familiarisation and test trials. For familiarisation trials, a 2 (Condition: AV vs. AO) by 2

(Block: 1 vs. 2) mixed-measures ANOVA was conducted with attention during familiarisation as the dependent variable. The main effects of Condition and Block were both significant (Condition: $F(1, 70) = 11.71$, $p = .001$, $\eta_p^2 = .14$; Block: $F(1, 70) = 10.89$, $p = .002$, $\eta_p^2 = .13$), while Condition x Block interaction was not ($F(1, 70) = 0.77$, $p = .38$, $\eta_p^2 = .01$). Infants in the AV condition attended more to the screen during familiarisation than their counterparts in the AO condition and both groups of infants attended more in Block 1 than in Block 2. As there was a main effect of Block, this Block 1 / Block 2 variable will be included in all subsequent analyses.

4.2.2 Test Trials: Word Segmentation Performance

Word segmentation performance was quantified by the difference in attention (that is, the overall proportion of looks to the screen) to Target versus Non-Target trials. Each participant was given a difference score (proportion looking to Target minus Non-Target trials), *d*, calculated for each block: larger difference scores index stronger preferences for Target trials, hence indicating better word segmentation performance. The *d* scores also account for the greater attention to the screen during familiarisation by infants in the AV condition compared to infants in the AO condition. Next, a 2 (Condition: AV vs. AO) x 2 (Block: 1 vs. 2) mixed-measures ANOVA was conducted with *d* scores as the dependent variable. The main effects of Condition and Block, and the Condition x Block interaction were not significant (Condition: $F(1, 70) = 1.57$, $p = .21$, $\eta_p^2 = .02$; Block: $F(1, 70) = 0.18$, $p = .68$, , $\eta_p^2 = .003$; Condition x Block: $F(1, 70) = 1.13$, $p = .29$, $\eta_p^2 = .016$; see Table 6 for means and SDs). These findings suggest that infants in AO and AV had similar word segmentation performance.

To examine whether there was successful word segmentation within each group of infants, *d* scores for each group were compared against zero via one-sample *t*-tests. *d* scores were significantly greater than zero in Block 1 but not Block 2 for infants in the AO

condition (Block 1: $t(17) = 2.68$, $p = .016$, Hedge's $g = 0.60$; Block 2: $t(17) = 0.76$, $p = .46$, Hedge's $g = 0.17$), whereas d scores were significantly greater than zero in both blocks for infants in the AV condition (Block 1: $t(18) = 2.51$, $p = .022$, Hedge's $g = 0.55$; Block 2: $t(18) = 2.49$, $p = .023$, Hedge's $g = 0.55$).

Table 5

Means (and Standard Deviations) of Attention (Proportion of Looks to Screen) during Familiarisation, Target and Non-Target Trials

	AO (N = 18)				AV (N = 19)			
	Familiar-isation	**Target**	**Non-Target**	***d* scores**	**Familiar-isation**	**Target**	**Non-Target**	***d* scores**
Overall	.61 (.19)	.61 (.15)	.57 (.13)	.04 (.08)	.76 (.16)	.68 (.18)	.61 (.15)	.07 (.10)
Block 1	.68 (.19)	.66 (.17)	.61 (.15)	.05 (.08)	.80 (.17)	.70 (.20)	.65 (.17)	.05 (.09)
Block 2	.48 (.23)	.52 (.17)	.50 (.19)	.03 (.16)	.68 (.20)	.62 (.19)	.52 (.20)	.10 (.17)

Note. *d* scores = Target – Non-target.

4.2.3 Does Gaze Behaviour Differ as a Function of Condition?

Separate Condition (AO vs. AV) x AOI Type (Eyes vs. Mouth) x Block (Block 1 vs. Block 2) mixed-measures ANOVAs for each trial type (Familiarisation, Target and Non-Target) were conducted to examine whether gaze behaviour differed between conditions for each trial type. None of the main effects or interactions were significant for any trial type (all Fs < 2.07, all ps > .15; see Table 6 for means and standard deviations, and Tables 7-9 for detailed ANOVA results). These results suggest that infants in both conditions did not attend differentially to the speaker's eyes versus mouth for Familiarisation, Target or Non-Target trials[2].

4.2.4 Do Individual Differences in Gaze Behaviour Influence Word Segmentation?

To investigate whether individual differences in gaze behaviour influence word segmentation, a hierarchical linear regression analysis was conducted with Attention (proportion of looks to screen during familiarisation), Condition, and Analysis Block as predictor variables, and segmentation performance (d scores) as the outcome variable. When Attention was entered in the first step of the regression analysis, the model was not significant ($F(1, 72) = 1.22$, $p = .27$, $\eta_p^2 = .02$), indicating that the variance in segmentation performance cannot be explained by attention to the screen during familiarisation ($\beta = .13$ [-0.06, 0.21], $p = .27$). To evaluate whether Condition and Analysis Block moderated the

[2] A Condition (AO vs. AV) x AOI Type (Eyes vs. Mouth) x Block (Block 1 vs. Block 2) x Trial Type (Target vs. Non-Target) mixed-measures ANOVA was conducted to examine if looks to the eye and mouth regions differed between Target and Non-Target trials as a function of condition. Specifically, the Condition x AOI Type x Trial Type interaction was of interest. Only the main effect of Condition was significant: Condition: $F(1, 280) = 3.88$, $p = .05$, $\eta_p^2 = .01$. Infants in the AV condition (mean = 0.31, SD = 0.25) paid greater attention to the eye and mouth regions than infants in the AO condition (mean = 0.25, SD = 0.26).
All other main effects and interactions were not significant: AOI Type: $F(1, 280) = 0.39$, $p = .53$, $\eta_p^2 = .001$; Block: $F(1, 280) = 0.37$, $p = .54$; Trial Type: $F(1, 280) = 0.67$, $p = .41$, Condition x AOI Type: $F(1, 280) = 2.89$, $p = .09$; Condition x Block: $F(1, 280) = 0.01$, $p = .93$, $\eta_p^2 < .001$; Condition x Trial Type: $F(1, 280) = 0.002$, $p = .96$, $\eta_p^2 < .001$; AOI Type x Block: $F(1, 280) = 0.41$, $p = .52$, $\eta_p^2 = .001$; AOI Type x Trial Type: $F(1, 280) = 0.23$, $p = .63$, $\eta_p^2 < .001$; Block x Trial Type: $F(1, 280) = 0.48$, $p = .49$, $\eta_p^2 = .001$; Condition x AOI Type x Trial Type: $F(1, 280) = 0.09$, $p = .76$, $\eta_p^2 < .001$; Condition x AOI Type x Block x Trial Type: $F(1, 280) = 0.06$, $p = .81$, $\eta_p^2 < .001$.

relationship between attention during familiarisation and segmentation performance, the Condition x Analysis Block interaction term was entered into the model in the second step of the regression analysis. The addition of this interaction term to the model was not significant ($\Delta R^2 = .03$, $\Delta F\ (3, 69) = 0.77$, $p = .52$, $\eta_p^2 = .03$), suggesting that Condition, Analysis Block and the Condition x Analysis Block interaction did not moderate the relationship between attention to the screen during familiarisation and segmentation performance (see Table 10 for the results).

To investigate whether individual differences in attention to the mouth (vs. the eyes) influence word segmentation performance, a mouth preference score for familiarisation was first derived by calculating the proportion of time spent looking at the mouth compared to the total amount of time spent looking at the mouth and eye regions during familiarisation trials, i.e., $mouth\ preference = \frac{attention\ to\ mouth}{(attention\ to\ mouth + attention\ to\ eyes)}$. Next, a hierarchical regression analysis was conducted with mouth preference during familiarisation and block as predictor variables and segmentation performance (*d* scores) as the outcome variable. This analysis only included data from infants in the AV condition. When mouth preference during familiarisation was entered in the first step of the regression analysis, the model was not significant ($F(1, 35) = 0.28$, $p = .60$, $\eta_p^2 = .008$), indicating that the variance in word segmentation performance of infants in the AV condition cannot be explained by their mouth preference during familiarisation. To examine whether the relationship between segmentation performance and mouth preference was moderated by the analysis block, Analysis Block and Mouth Preference x Analysis Block interaction term were entered into the model in the second step of this regression analysis. The addition of these variables was not significant ($\Delta R^2 = .04$, $\Delta F\ (2, 35) = 0.62$, $p = .54$, $\eta_p^2 = .03$), suggesting that Analysis Block and Mouth Preference x Analysis Block did not moderate the relationship between segmentation performance and mouth preference during familiarisation (see Table 11 for the results).

Table 6

Means (and Standard Deviations) of Attention to the Speaker's Eye and Mouth Region Across Trial Types for Each Condition

	AO			AV		
	Familiarisation	**Target**	**Non-Target**	**Familiarisation**	**Target**	**Non-Target**
Overall						
Eyes	0.30 (0.26)	0.31 (0.25)	0.27 (0.26)	0.34 (0.26)	0.32 (0.24)	0.28 (0.23)
Mouth	0.22 (0.25)	0.23 (0.25)	0.21 (0.25)	0.30 (0.27)	0.33 (0.27)	0.33 (0.26)
Block 1						
Eyes	0.35 (0.28)	0.30 (0.27)	0.31 (0.28)	0.35 (0.26)	0.33 (0.25)	0.30 (0.24)
Mouth	0.22 (0.27)	0.22 (0.26)	0.21 (0.27)	0.30 (0.27)	0.32 (0.29)	0.34 (0.28)
Block 2						
Eyes	0.25 (0.23)	0.31 (0.24)	0.24 (0.25)	0.34 (0.28)	0.31 (0.24)	0.25 (0.22)
Mouth	0.23 (0.25)	0.24 (0.25)	0.21 (0.25)	0.30 (0.29)	0.33 (0.26)	0.32 (0.24)

Table 7

Results of Condition (AO vs. AV) x AOI (Eye vs. Mouth) x Block (Block 1 vs. Block 2) Mixed-Measures ANOVAs for Familiarisation Trials

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*
(Intercept)	9.81	140	0.07		
Condition	0.13	1	0.13	1.92	.17
AOI	0.13	1	0.13	1.87	.17
Block	0.02	1	0.02	0.27	.60
Condition x AOI	0.01	1	0.01	0.14	.71
Condition x Block	0.02	1	0.02	0.24	.62
AOI x Block	0.03	1	0.03	0.43	.51
Condition x AOI x Block	0.02	1	0.02	0.31	.58

Table 8

Results of Condition (AO vs. AV) x AOI (Eye vs. Mouth) x Block (Block 1 vs. Block 2) Mixed-Measures ANOVAs for Target Trials

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*
(Intercept)	9.31	140	0.07		
Condition	0.12	1	0.12	1.82	.18
AOI	0.04	1	0.04	0.59	.44
Block	0.0002	1	0.0002	0.004	.95
Condition x AOI	0.06	1	0.06	0.96	.33
Condition x Block	0.002	1	0.002	0.04	.85
AOI x Block	0.005	1	0.005	0.07	.80
Condition x AOI x Block	0.001	1	0.001	0.02	.90

Table 9

Results of Condition (AO vs. AV) x AOI (Mouth vs. Eyes) x Block (Block 1 vs. Block 2) Mixed-Measures ANOVAs for Non-Target Trials

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*
(Intercept)	8.97	140	0.06		
Condition	0.13	1	0.13	2.07	.15
AOI	0.0007	1	0.0007	0.01	.92
Block	0.06	1	0.06	0.86	.36
Condition x AOI	0.13	1	0.13	2.04	.16
Condition x Block	0.0002	1	0.0002	0.004	.95
AOI x Block	0.03	1	0.03	0.42	.52
Condition x AOI x Block	0.003	1	0.003	0.05	.83

Table 10

Summary of Hierarchical Regression Analysis Predicting Segmentation Performance

Predictors	β	95% CI [lower, upper]	t	F	p	R^2	ΔR^2
Step One				1.22	.27	.02	.02
Attention during Familiarisation	.13	[-0.06, 0.21]	1.11		.27		
Step Two				0.88	.48	.05	.03
Attention during Familiarisation	.11	[-0.09,0.22]	0.81		.42		
Condition	-.006	[-0.09, 0.09]	-0.04		.97		
Block	-.03	[-0.10, 0.09]	-0.17		.86		
Condition x Block	.10	[-0.06, 0.18]	0.97		.34		

Table 11

Summary of Hierarchical Regression Analysis of Mouth Preference and Segmentation Performance (Auditory-Visual Condition Only)

Predictors	β	95% CI [lower, upper]	t	F	p	R^2	ΔR^2
Step One				0.28	.60	.008	.008
Mouth Preference (Familiarisation)	.09	[-0.10, 0.17]	0.53		.60		
Step Two				0.51	.68	.04	.04
Mouth Preference (Familiarisation)	.12	[-0.16, 0.26]	0.49		.63		
Block	.22	[-0.10, 0.22]	0.80		.43		
Mouth Preference (Familiarisation) x Block	-.05	[-0.30, 0.26]	-0.15		.88		

4.2.5 Time Course Analyses

The lack of a significant relationship between word segmentation and gaze behavior was unexpected given that it has been previously established that infants attend to the eye and mouth regions differentially depending on whether the face is producing speech (Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013). One possible explanation for these unexpected findings is that word segmentation performance here is indexed by difference scores which were derived from PTLs averaged across time, which may mask any potential temporal variations in looking behavior. To explore this possibility, the time course of infants' looking behavior was examined. Gaze data were aggregated into 20 time-bins (~1.24s per bin for Familiarisation trials, and ~0.80s per bin for Target and Non-Target trials) and compared sequentially using ANOVAs to identify any particular time period during which infants' looking patterns to the face, eye and mouth AOIs diverged. The false discovery rate (FDR) method (Benjamini & Yekutieli, 2001) was used to correct for multiple comparisons. For these analyses, PTLs to the face, eyes, and mouth AOIs were derived from equations (1-3) (see 4.1.5 Eye Tracking Analyses) in which total fixation duration serves as the denominator.

First, infants' attention during Familiarisation, Target and Non-Target trials was examined. Figures 17, 18, and 19 illustrate time-courses for eye AOI, mouth AOI and face AOI, respectively. The separate time course analyses for attention to the face and for attention the eyes indicated that there were no significant differences between the AO and the AV groups. The only significant differences between the two groups of infants were for attention to the mouth. As can be seen in Figure 19, attention to the mouth region was generally greater in the AV than in the AO condition early in the time course but this difference gradually declined over time. In Block 1, the greater attention to the mouth in the AV over the AO group was significant only for Non-Target trials and only from 0.00 to 6.42s

(*p*s < .05 for all time bins), but not after that time. For Block 2, the initial AV > AO difference in attention to the mouth region was significant from 0.00-5.62s (*p*s < .05 for all time bins) for Non-Target trials, and for Target trials from 0.00-5.63s (*p*s < .02 for all time bins).

Next, given that the only significant effect was for attention to the mouth region, separate time-course analyses were conducted for each group to examine attention to mouth for Target versus Non-Target trials over time (Figure 20). Neither of these analyses showed any significant results (all *p*s > .41), suggesting that infants did not attend to the mouth differentially for Target versus Non-Target trials.

Figure 17

Time Courses of Attention to the Face AOI for Each Trial Type

Figure 18

Time Courses of Attention to the Eye AOI for Each Trial Type

Figure 19

Time Courses of Attention to the Mouth AOI for Each Trial Type

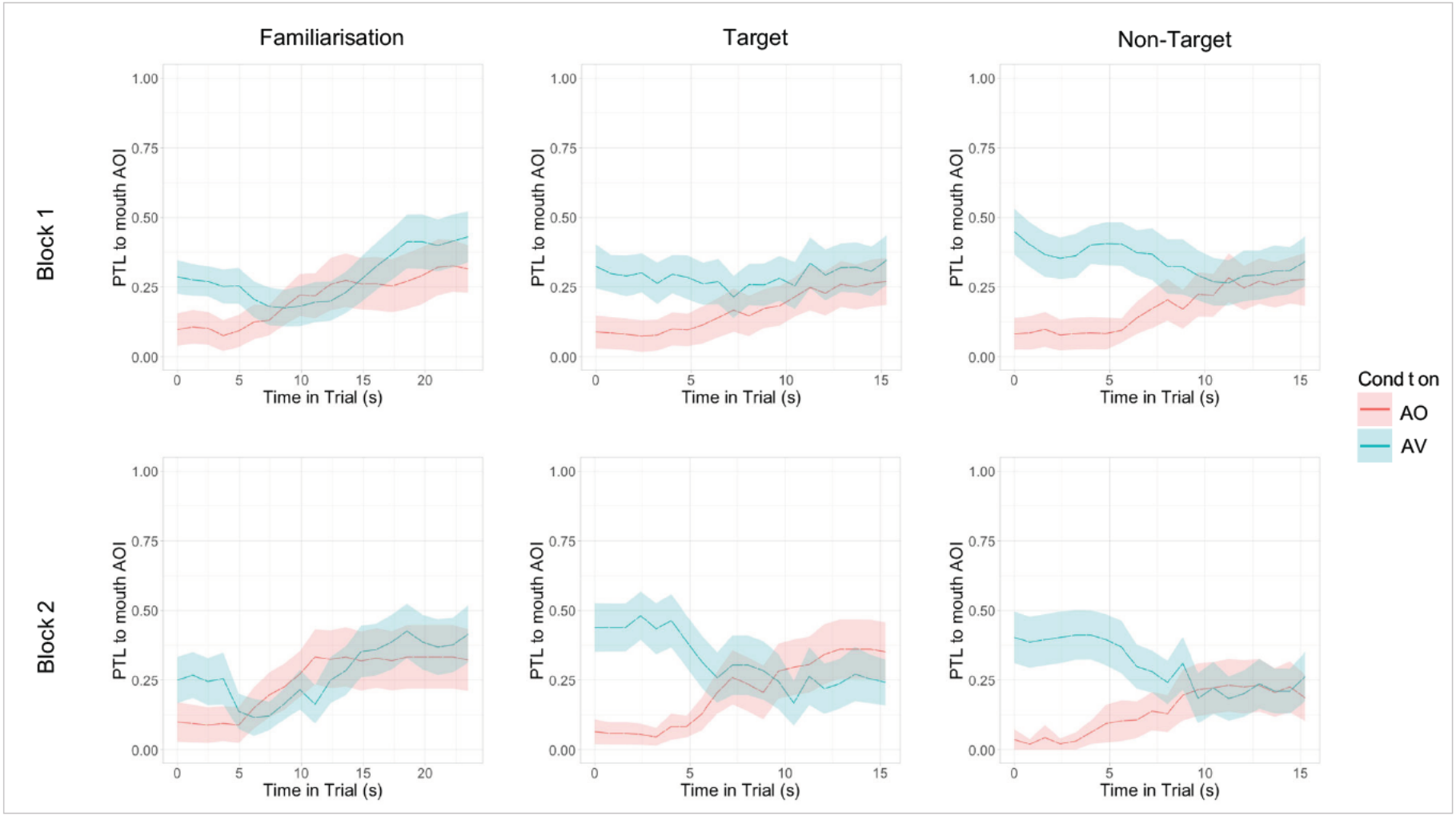
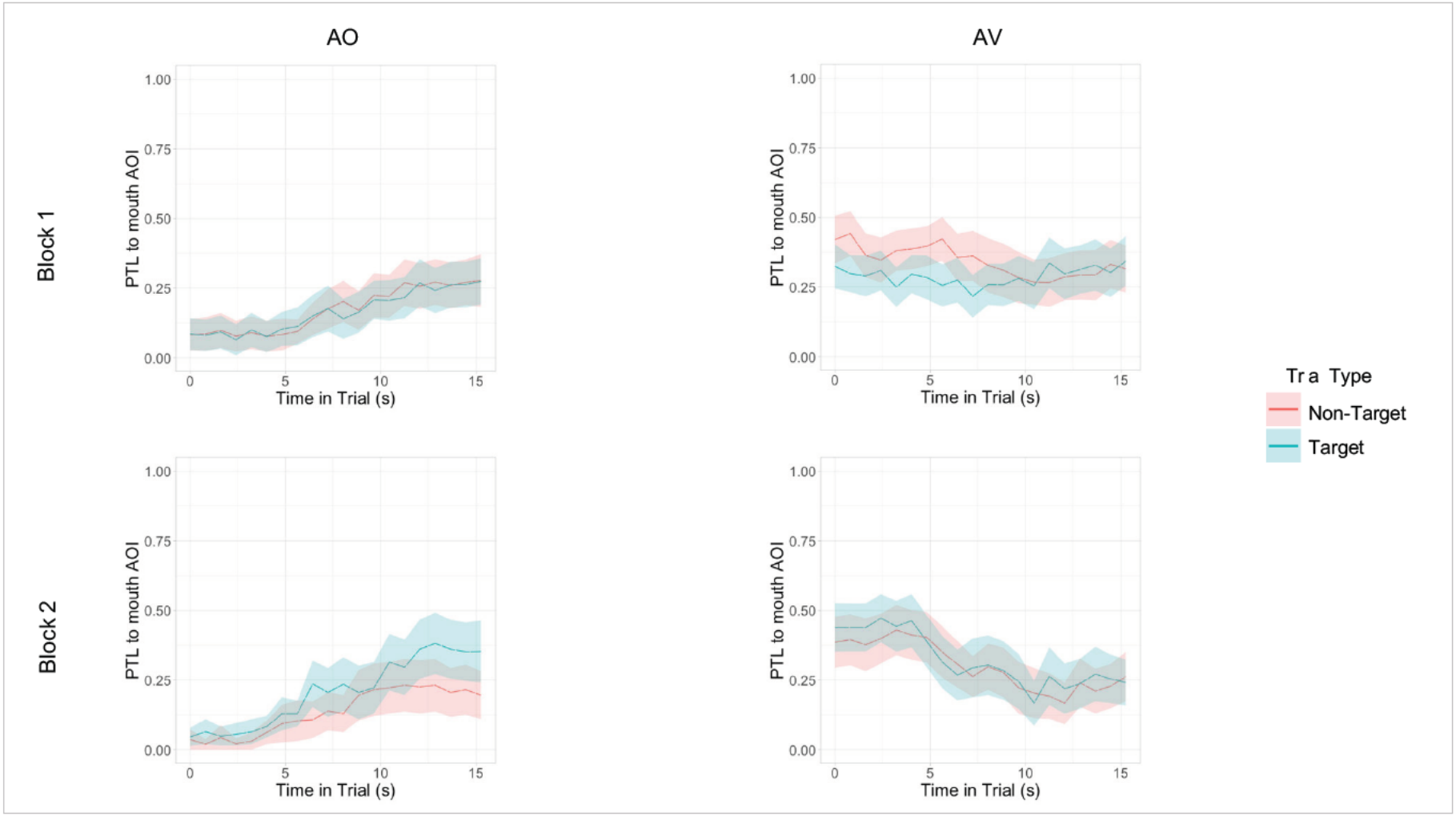


Figure 20

Time Courses of Attention to the Mouth AOI Between Target and Non-Target Trials for Each Condition



4.3 Discussion

To examine whether visual speech information augments infant word segmentation performance, two groups of 7.5-month-old infants were tested. One group was presented with a static image of the speaker's face paired with auditory recordings of Familiarisation passages, then Target and Non-Target word tokens, while the second group of infants was presented with dynamic videos of the speaker's talking face reciting the passages and word tokens, then Target and Non-Target word tokens. Preliminary analyses of attention during familiarisation revealed that infants in the auditory-visual condition attended more to the screen than infants in the auditory-only condition, and that attention was greater during the first than second block.

To quantify segmentation performance, a difference score was calculated by subtracting attention to Non-Target word tokens from attention to Target word tokens. If difference scores are significantly greater than zero, then word segmentation is successful. The second set of analyses showed that there was successful word segmentation in the first block for infants in both the AO and the AV condition, and in the second block for only infants in the AV condition[3], suggesting that successful word segmentation was present in both conditions, but sustained only for infants in the AV condition.

The next set of analyses that investigated whether gaze behavior influences word segmentation revealed a number of unexpected findings. First, both groups of infants attended similarly to the speaker's eye and mouth regions in all three types of trials (Familiarisation, Target and Non-Target). Second, attention to the screen during

[3] To control for greater attention to the screen by the AV group in general, a Condition (AO vs. AV) x Block (Block 1 vs. Block 2) mixed-measures ANOVA was conducted using $\frac{attention\ to\ Target}{attention\ to\ Target + attention\ to\ Non\ \ Target}$ as the dependent variable. The main effects of Condition and Block and the Condition x Block interaction were not significant: Condition: $F(1, 70) = 0.70$, $p = .41$, $\eta_p^2 = .01$; Block: $F(1, 70) = 1.16$, $p = .29$, $\eta_p^2 = .02$; Condition x Block: $F(1, 70) = 0.69$, $p = .41$, $\eta_p^2 = .01$. Successful word segmentation in the AV group cannot be explained by greater fixation durations to the screen during Target and Non-Target trials.

familiarisation did not predict word segmentation performance, and this relationship was not moderated by condition or block. This finding suggests that successful word segmentation performance in Block 2 by infants in the AV condition cannot be entirely explained by their greater attention to the screen during familiarisation. Third, mouth preference did not predict word segmentation performance for infants in the AV condition. These results were surprising because previous studies have found that infants attend more to the mouth region than the eye region of a talking face and this shift in attention increases with age (e.g., Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013) and it has been postulated that infants are able to make use of visual (e.g., mouth movements) and auditory modalities to segment speech (Kitamura et al., 2014). One possibility for the null findings is that these effects may have been masked by the use of PTLs which are averaged across time.

The final set of analyses overcame this potential problem by investigating the time course of infants' looking behaviour. There was no group difference in looking behavior to the face over time, indicating that infants' attention to the face was similar regardless of whether a static image or dynamic talking face was presented to them. When gaze patterns to the eye and mouth regions were analysed specifically, there were AV vs. AO group differences only for the mouth and not for the eye region. Infants in the AV condition attended more to the mouth region than infants in the AO condition during the first 5s of Non-Target trials in both Blocks 1 and 2, and during the first 5s of Target trials but only in Block 2. Time course plots showed a general decrease in attention to the mouth over time for the AV group, and an increase in attention to the mouth over time for the AO group, suggesting that infants in the AO condition may be sourcing linguistic information from the speaker's face. Interestingly, attention to the mouth AOI over time did not differ between Target and Non-Target trials even for infants in the AV condition, suggesting that infants' looking behaviour to the mouth region was not influenced by the 'novelty' of the words.

The finding that the visual speech benefit was observed only in the second block was somewhat unexpected. However, when coupled with the finding that attention does not predict segmentation performance even for the AV group of infants, it narrows the scope of potential pathways via which visual speech information augments infant speech perception. First, visual speech cues may provide an additional modality through which information is processed. As phonological short-term memory is shown to be fundamental for successful word segmentation (Minagawa et al., 2017) and in accordance with the dual coding theory of memory (Paivio, 1991), it stands to reason that visual speech cues strengthen the memory trace of the primitive phonological representation of newly-segmented words, resulting in the sustained segmentation found in the second block only for the AV group. Second, as the videos presented to the AV group involved a close-up of the speaker talking in IDS, infants in the AV group may have inferred socially meaningful cues (e.g. eye contact) which may have inadvertently enhanced their segmentation performance. This is in line with neurophysiological findings that infants are already sensitive to ostensive signals such as directed gaze and IDS by 5 months of age (Parise & Csibra, 2013), and behavioural findings that communicative social contexts such as contingent responding (e.g, Goldstein & Schwade, 2008; Goldstein et al., 2010; Mackensen & Grossmann, 2015) and gestures (Yoon et al., 2008) foster learning in infants. During the experimental session for the present study, infants in the AV group tended to display communicative behaviour such as reaching for the screen and babbling in response to the video presentations. Third, visual cues from the speaker's talking face may provide intersensory redundancies that support language perception (Bahrick & Lickliter, 2000; Gogate & Bahrick, 1998; 2001; Gogate et al., 2006) as IDS is associated with exaggerated facial cues (Green et al., 2010) that are temporally synchronised with the auditory prosody exaggerations in IDS (Shepard et al., 2012). These, in combination with the increased attention that likely stems from the exaggerated facial

expressions produced alongside IDS, may account for the persistence of segmentation that was found only for the AV group. Given the important facilitatory role that IDS plays in word segmentation (Thiessen et al., 2005) and word learning (e.g., Graf Estes & Hurley, 2013), examining whether the visual speech benefit found in this study is also evident when ADS is presented may further clarify the role of visual speech cues in infant language perception.

An alternative explanation for this finding is that the AO group were transitioning from a familiarity preference to a novelty preference after being repeatedly exposed to the stimulus (Burnham & Dodd, 1999; Houston-Price & Nakai, 2004) resulting in the lack of a significant difference observed in block 2. This is unlikely because the AV group of infants continued to show a familiarity preference even though they attended more to the screen than the AO group of infants in the first block. Another counterargument then would be that attention to the screen may not be a good index of attention for the AO group of infants in general because still photographs are less appealing than dynamic videos, and thus the AO group of infants may still have been attending to the stimulus even though they may not have been looking at the screen. To address this possibility, future work could implement control conditions by presenting videos during familiarisation and still photographs during test phases and vice versa.

It is instructive to compare these results with those of Hollich et al. (2005). In that study, the auditory recordings were always paired with noise and 7.5-month-olds successfully segmented speech only in the AV, but not in the AO condition. Here, with no auditory noise, word segmentation was found in both AV *and* AO conditions, and successful word segmentation persisted into the second block only in the AV condition. It may be that infants can segment clear speech effectively by 7 months and adding visual speech information does not provide substantial incremental benefit.

To investigate whether this is the case, future work could examine if visual speech benefit is more pronounced in younger infants, i.e., they may segment AV but not AO speech, or at least have greatly reduced segmentation for AO speech. Such a result is likely because infants younger than 7 months can segment words under certain conditions. For example, 6-month-olds can segment speech when the target word is placed next to their own name or a highly familiar word (Bortfeld et al., 2005). Additionally, 6-month-olds can recognise familiar words (Bergelson & Swingley, 2012), indicating that they have already isolated these words from the speech stream. Given these findings, coupled with the finding that newborns can integrate visual and auditory information (Guellaï et al., 2016), it may even be the case that 6-month-olds will show an AV over AO advantage right from the first block.

Looking behaviour to the eye and mouth regions was not associated with the amount of visual speech benefit derived in this task. This pattern of results differs from studies with infants that have found a positive relationship between looking behaviour to the speaker's mouth and language development (e.g., Elsabbagh et al., 2013; Young et al., 2009). Additionally, looking behaviour was surprisingly not different for target and non-target words in the present study. It was hypothesised that infants would attend to the mouth longer in non-target test trials than in target test trials because infants were not exposed to the non-target words in the familiarisation phase and because it has been previously suggested that infants direct their attention to the mouth to gather linguistic information (e.g., Tenenbaum et al., 2013). A likely explanation for the non-significant difference between looking behaviour during target and non-target trials (in PTL and time-course analyses) is that the non-target words were not entirely novel— *cup*, *dog*, *bike*, and *feet* are words that infants commonly hear in their everyday lives. Further work on looking behaviour using novel non-words would expound whether this is indeed the case.

Word segmentation is arguably one of the most important skills that infants must master before acquiring a language. Most infant segmentation studies have employed auditory-only stimuli despite the multimodality of speech. Studying the role of visual speech cues in infant language development has important implications particularly for infants who do not have access to a clear auditory signal. This study found that 7.5-month-olds derive some form of benefit by having access to the speaker's talking face even though auditory cues are sufficient for them to successfully segment continuous speech, and that individual differences in looking behaviour to the speaker's eye and mouth regions were not related to segmentation performance. Critically, there is still much to learn about the mechanisms via which visual speech information enhances speech perception in infancy.

CHAPTER 5

A Neurophysiological-Behavioural Exploration of Auditory-Visual Speech Perception

Data from behavioural studies are typically more readily interpretable than those from neurophysiological studies, however, as neurophysiological studies do not require overt responses and are less susceptible to task demands, they are often used to corroborate behavioural findings or even point to mechanisms that are not manifested at the behavioural level. In Study 1, visual speech benefit was examined at the neural level: cortical tracking of the speech envelope was compared across auditory-only, visual-only and auditory-visual conditions with visual speech benefit indexed by the additive model criterion. Cortical tracking of the speech envelope was generally found to be better during auditory-visual speech compared to visual-only or auditory-only speech across ages (5-month-olds, 4-year-olds, and adults). In Study 2, 7.5-month-old infants' visual speech benefit was examined at the behavioural level through word segmentation: more robust word segmentation was found in infants in the auditory-visual condition compared to infants in the auditory-only condition (and this was indexed solely by persistent word segmentation into the second block in the auditory-visual condition).

The aim of this chapter is to explore the interactions between neurophysiological and behavioural indices of auditory-visual speech perception. Study 1 demonstrated that visual speech information enhances cortical tracking of the speech envelope. To shed light on the exact mechanism(s) underlying the observed facilitatory effects of visual speech information, two research questions are addressed here:

1. Is looking behaviour to a talker's face related to simultaneous cortical tracking of the speech envelope by 5-month-olds, 4-year-olds and adults?

2. Is cortical tracking at 5 months of age related to subsequent word segmentation performance at 7 months?

To answer these research questions, cortical tracking was quantified by stimulus reconstruction accuracy here for the following reason. In Study 1, both stimulus reconstruction accuracy of decoders (backward modelling) and prediction accuracy of temporal response functions (TRFs) (forward modelling) were used to investigate envelope tracking. The pattern of results observed in Study 1 was similar when stimulus reconstruction accuracy and prediction accuracy were examined, with the only key difference being that a visual speech benefit [i.e., AV > (A+V)] was observed in 5-month-olds and adults for prediction accuracy but not for stimulus reconstruction accuracy. In addition, stimulus reconstruction accuracy is a more sensitive measure than the TRFs because it maps EEG data from all channel simultaneously. To avoid duplicating analyses, only the more sensitive measure of cortical tracking—stimulus reconstruction accuracy—was used in the current investigation.

5.1 Part I: The Relationship between Looking Behaviour and Auditory-Visual Speech Perception

To reap the benefits that visual speech cues bring to speech perception, individuals must actively attend to a speaker's talking face. Two main facial regions to which gaze is often directed are the eyes and mouth: 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). Evidence from face viewing studies indicate that humans are cognizant 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 & McConkie, 1999). This attentional shift is observed even in infants as young as 6 months (Tenenbaum et al., 2013). Examinations of looking behaviour have

generally been conducted at the group level even though systematic individual differences exist. A seminal study by Yarbus (1967) illustrated the individual variations in looking behaviour—eye scanning movements to a static face typically centered on the eye and mouth regions but no two individuals showed the exact same scanning pattern. This typical facial scanning pattern focusing on the eye and mouth regions has been replicated (Arizpe et al., 2012; Blais et al., 2008; Peterson & Eckstein, 2012) and extended to dynamic faces (Gurler et al., 2015; Mehoudar et al., 2014) with persistent interindividual differences. More importantly, these idiosyncratic differences are related to perceptual performance (Gurler et al., 2015; Mehoudar et al., 2014; Peterson & Eckstein, 2012). Emotion recognition, gender and identity discrimination are significantly reduced when individuals' personal preferred fixation points are disrupted (Peterson & Eckstein, 2012), and individuals who report experiencing the McGurk effect more frequently also spend a larger proportion of time fixating on the speaker's mouth (Gurler et al., 2015). Altogether, these findings point toward the strong likelihood that individuals' idiosyncratic preferences in the 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 correspond 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. 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 to 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 proposed 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 & Hansen-Tift, 2012). Adults, by comparison, 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 (Vatikiotis-Bateson et al., 1998).

Face viewing and speech perception studies suggest the possibility that individuals' idiosyncratic differences in looking patterns to a talker's face will influence the extent of benefit that visual speech cues bring to speech perception. Here this possibility is investigated by inspecting the gaze and EEG data that were simultaneously recorded for the three age groups (5-month-olds, 4-year-olds, and adults) in Study 1. Accordingly, fixation durations to the speaker's face and its relationship with cortical tracking of the speech envelope were examined. Five-month-olds were expected to show a different pattern of results from 4-year-olds and adults. As 5-month-olds are likely to be in the process of shifting their attentional focus from the speaker's eyes to the speaker's mouth region, it is hypothesised that the proportion of time spent attending to the speaker's mouth will be positively correlated with cortical tracking in VO and AV conditions for 5-month-olds. Moreover, 5-month-olds are in the process of acquiring language, and any additional information that can be extracted from visual speech cues may aid language acquisition. The same positive correlation is not expected for 4-year-olds and adults as older infants and adults have been found to focus more

on the speaker's eyes when the auditory speech signal is clear (e.g., Lewkowicz & 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. Instead, as the speaker's face occupies more than 50% of the screen, overall attention to the screen may more accurately reflect the amount of information that 4-year-olds and adults extract from a talking face. Thus, in contrast to the prediction of a mouth region correlation with cortical tracking for 5-month-olds, for 4-year-olds and adults it is expected that overall attention to the screen will be positively correlated to cortical tracking when visual speech information is available (VO and AV conditions).

5.1.1 Methods

5.1.1.1 Participants.

Data from all participants in Study 1 (18 five-month-olds, 19 four-year-olds, 18 adults) were included in the analyses here.

5.1.1.2 Stimuli, Apparatus, and Procedure.

Stimuli, apparatus and procedure are described in Study 1 (Section 3.1). As mentioned in Study 1 (Section 3.1.3), prior to the onset of stimulus presentation, gaze data were collected at the four corners and centre of the screen to calibrate the individual's eye movements and allow computation of any spatial offsets in gaze data.

Means and standard deviations of the spatial offsets (x- and y-coordinates) for each age group are reported in Table 12. 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 due to noisy EEG data, videos of participants who met the EEG data inclusion criterion (≤ 20 noisy channels) in Study 1 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.

5.1.1.3 Gaze Measures.

Areas of interest (AOIs) covering the top half and bottom half of the speaker's face demarcated the speaker's eye and mouth regions (Figure 21). These AOIs were of equal dimensions (640 x 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 = $[\frac{total\ fixation\ duration}{trial\ duration}]$, (hereafter referred to as *Attention*) and

(2) Proportion looking to the speaker's mouth region (hereafter referred to as *PTL Mouth*) = $[\frac{total\ fixation\ duration\ to\ mouth}{total\ fixation\ duration\ to\ mouth + total\ fixation\ duration\ to\ eyes}]$.

Note that PTL Mouth is a relative measure of mouth compared to eyes, so chance is 0.5, scores > .5 show greater fixation to mouth than eyes and scores < .5 show greater fixation to eyes than mouth. All statistical analyses on these two gaze measures were conducted using custom scripts in MATLAB R2019a (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.

Table 12

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)

Figure 21

Areas of Interest (AOIs) Defined for the Speaker's Eye and Mouth Regions

5.1.2 Results

Data were analysed in two parts. First, ANOVAs were conducted for each age group to examine the differences in attention and proportion looking to speaker's mouth between conditions. All means and standard deviations of attention and proportion looking to speaker's mouth are reported in Table 13. Next, correlations were conducted for each condition to examine the relationship between (1) cortical tracking and attention, and (2) cortical tracking and looking preference for each age group, where cortical tracking is quantified by stimulus reconstruction accuracy.

5.1.2.1 Gaze Behavior: Attention.

To examine whether attention differed between conditions, separate one-way within-subjects ANOVAs were conducted for each age group with Attention as the dependent variable and Condition as the independent variable. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for 4-year-olds, $\chi^2(2) = 8.41$, $p = .01$, thus the Greenhouse-Geisser correction was applied to this age group. The analyses revealed a significant main effect of Condition for all age groups (5-month-olds: $F(2, 34) = 3.58$, $p = .04$, $\eta_p^2 = .17$; 4-year-olds: $F(1.44, 25.89) = 26.67$, $p < .001$, $\eta_p^2 = .60$; adults: $F(2, 34) = 7.16$, $p = .002$, $\eta_p^2 = .30$). Subsequent post-hoc comparisons between conditions were made using paired-sample *t*-tests.

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

Four-year-olds: Attention was significantly greater in the AV than in the AO ($t(18) = 6.10$, $p < .001$, Hedge's $g = 1.54$) and in the VO condition ($t(18) = 9.19$, $p < .001$, Hedge's g

= 1.43), whereas the difference in attention between AO and VO conditions was not significant ($t(18) = -1.00$, $p = .33$, Hedge's $g = -0.26$).

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

5.1.2.2 Gaze Behaviour: PTL Mouth.

To examine whether looking behaviour to the speaker's face differed between conditions, separate one-way within-subjects ANOVAs were conducted for each age group with PTL Mouth as the dependent variable and Condition as the independent variable. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for adults, $\chi^2(2) = 10.40$, $p = .006$, therefore the Greenhouse-Geisser correction was applied to that age group. The one-way ANOVAs were significant for 5-month-olds and adults (5-month-olds: $F(2, 26) = 4.98$, $p = .01$, $\eta_p^2 = .28$; adults: $F(1.35, 23.00) = 13.40$, $p < .001$, $\eta_p^2 = .44$), but not for 4-year-olds ($F(2, 14) = 1.82$, $p = .20$, $\eta_p^2 = .21$). Subsequent analyses involved one-sample *t*-tests to assess whether PTL Mouth was significantly greater than chance and paired-sample *t*-tests 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: $t(14) = -.72$, $p = .48$, Hedge's $g = -0.18$; VO: $t(13) = .19$, $p = .85$, Hedge's $g = 0.05$; AV: $t(13) = .14$, $p = .89$, Hedge's $g = 0.10$). Next, paired-sample *t*-tests indicated that infants' looking preference for the speaker's mouth was greater in the VO than the AO condition ($t(13) = 3.44$, $p = .004$, Hedge's $g = 0.16$), and in the AV than the AO condition ($t(13) = 2.82$, $p =$

.015, Hedge's $g = 0.28$), but the difference between VO and AV conditions was not significant, $t(13) = 0.17$, $p = .87$, Hedge's $g = 0.01$.

Four-year-olds: One-sample *t*-tests indicated that PTL Mouth was not significantly greater than chance in any condition (AO: $t(7) = -.23$, $p = .83$, Hedge's $g = -0.07$; VO: $t(8) = .49$, $p = .63$, Hedge's $g = 0.14$; AV: $t(8) = -.09$, $p = .93$, Hedge's $g = 0.09$). 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: $t(7) = -1.80$, $p = .11$, Hedge's $g = -0.35$; AO vs. AV: $t(7) = -.65$, $p = .53$, Hedge's $g = -0.14$; VO vs. AV: $t(8) = 2.28$, $p = .05$, Hedge's $g = 0.13$).

Adults: One-sample *t*-tests indicated that PTL Mouth was significantly greater than chance in the VO condition ($t(17) = 4.93$, $p < .001$, Hedge's $g = 1.11$), but not in the AO or AV conditions (AO: $t(17) = .47$, $p = .64$, Hedge's $g = 0.11$; AV: $t(17) = 1.43$, $p = .17$, Hedge's $g = 0.32$). 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: $t(17) = -4.12$, $p < .001$, Hedge's $g = -0.81$; AO vs. AV: $t(17) = -2.57$, $p = .02$, Hedge's $g = -0.21$; VO vs. AV: $t(17) = 3.28$, $p = .004$, Hedge's $g = 0.58$).

Table 13

Means (and Standard Deviations) of Overall Attention and PTL to Speaker's Mouth Region (PTL Mouth) Across Ages

	5-month-olds		4-year-olds		Adults	
	n	**Mean (SD)**	***n***	**Mean (SD)**	***n***	**Mean (SD)**
AO						
Attention	18	56.88 (11.41)	19	55.00 (13.74)	18	76.57 (12.49)
PTL Mouth	15	43.84 (33.03)	7	47.99 (24.79)	18	53.38 (30.20)
VO						
Attention	18	53.37 (8.28)	19	58.36 (11.23)	18	81.73 (13.91)
PTL Mouth	14	51.65 (32.96)	8	54.63 (28.06)	18	75.52 (21.96)
AV						
Attention	18	58.47 (11.16)	19	74.60 (10.91)	18	81.55 (11.54)
PTL Mouth	14	51.20 (32.67)	8	49.04 (33.14)	18	59.92 (29.52)

5.1.2.3 Overall Attention and Cortical Tracking.

To investigate the relationship between attention and cortical tracking, Pearson's correlations between attention and stimulus reconstruction accuracy were conducted for each condition and age group. Interestingly, attention was not significantly correlated with stimulus reconstruction accuracy for any condition in any of the three age groups (all rs < .42, all ps > .08; see Table 14 for details).

5.1.2.4 PTL Mouth and Cortical Tracking.

Although overall attention was not significantly correlated with the strength of cortical tracking, individual differences in proportion looking time to the speaker's mouth region (vs. eye region) may still be associated with the strength of cortical tracking. Pearson's correlations between PTL Mouth and stimulus reconstruction accuracy were conducted for each condition and age group. The correlation between 5-month-olds' relative attention to the speaker's mouth region and stimulus reconstruction accuracy was significant for VO condition ($r(14) = .67$, $p = .009$) but not for AO or AV and not for any condition for the 4-year-olds or adults (all rs < .29, all ps > .19; see Table 14 for details).

Table 14

Summaries of Pearson's Correlations Between the Two Gaze Measures (Attention and PTL Mouth) and Stimulus Reconstruction Accuracy for All Age Groups

	5-month-olds			**4-year-olds**			**Adults**		
	n	***r***	***p***	**n**	***r***	***p***	**n**	***r***	***p***
AO									
Attention	18	.09	.73	19	-.01	.96	18	-.28	.25
PTL Mouth	15	.27	.33	8	.04	.93	18	-.32	.19
VO									
Attention	18	.16	.53	19	-.25	.31	18	-.07	.77
PTL Mouth	14	.67	**.009**	8	.13	.75	18	.02	.95
AV									
Attention	18	.42	.09	19	-.22	.36	18	-.01	.97
PTL Mouth	14	.29	.32	8	-.22	.58	18	-.18	.47

5.1.3 IDS vs. ADS

In Study 1, adult participants were additionally presented with ADS in AO, VO and AV conditions. Gaze data collected during that session were analysed as an exploratory investigation of (a) whether adults' gaze behaviour differed between speech types, and (b) whether the relationship between gaze behaviour and neural tracking differed between speech types.

5.1.3.1 Gaze Behaviour: Attention.

To investigate whether adults' gaze behaviour differed between IDS and ADS, separate Speech Type (IDS vs. ADS) x Condition (AO vs. VO vs. AV) repeated-measures ANOVAs were conducted for the two gaze measures (i.e., Attention, PTL Mouth). All means and standard deviations are reported in Table 15.

As Mauchly's test of sphericity was significant ($\chi^2(14) = 49.56$, $p < .001$), the Greenhouse-Geisser correction was applied. The main effect of Condition was significant ($F(0.42, 12.72) = 15.88$, $p < .001$, $\eta_p^2 = .51$), but the main effect of Speech Type and the Speech Type x Condition interaction were not significant (Speech Type: $F(0.42, 6.36) = 1.16$, $p = .30$, $\eta_p^2 = .07$; Speech Type x Condition: $F(0.85, 12.72) = 0.98$, $p = .39$, $\eta_p^2 = .06$). Even though the main effect of Speech Type and the Speech Type x Condition interaction were not significant, a one-way repeated-measures ANOVA was conducted for ADS with Attention as the dependent variable and Condition as the independent variable. The Greenhouse-Geisser correction was applied as Mauchly's test of sphericity was significant, $\chi^2(2) = 22.48$, $p < .001$. The one-way ANOVA revealed a significant main effect of Condition ($F(1.11, 16.67) = 12.31$, $p < .001$, $\eta_p^2 = .45$).

Post-hoc paired-sample *t*-tests (adjusted alpha level of .05/3) were conducted to examine the differences between conditions for ADS only. Adults attended to the screen longer in the VO than in the AO condition ($t(15) = 3.86$, $p = .005$, Hedge's $g = 0.58$), and in

the VO than the AV condition ($t(15) = 4.46$, $p = .001$, Hedge's $g = 0.18$), whereas the difference between AO and AV conditions was not significant ($t(15) = -2.90$, $p = .03$, Hedge's $g = -0.39$).

When ADS was compared against IDS, attentional differences between speech types were not significant in any condition (AO: $t(15) = 0.53$, $p = .60$, Hedge's $g = 0.14$; VO: $t(15) = 1.76$, $p = .10$, Hedge's $g = 0.32$; AV: $t(15) = 0.73$, $p = .48$, Hedge's $g = 0.13$).

5.1.3.2 Gaze Behaviour: PTL Mouth.

As Mauchly's test of sphericity was significant ($\chi^2(14) = 56.33$, $p < .001$), the Greenhouse-Geisser correction was applied. The main effect of Speech Type and the main effect of Condition were significant (Speech Type: $F(0.54, 8.02) = 7.16$, $p = .02$, $\eta_p^2 = .32$; Condition: $F(1.07, 16.05) = 17.95$, $p < .001$, $\eta_p^2 = .54$), while the Speech Type x Condition interaction was not ($F(1.07, 16.05) = 0.03$, $p = .97$, $\eta_p^2 = .002$).

To investigate whether proportion of time spent fixating on the speaker's mouth region (vs. eye region) was significantly greater than chance, one-sample *t*-tests were conducted for ADS only. These tests that adults' relative attention to the speaker's mouth region was not significantly greater than chance across conditions (AO: $t(15) = -1.57$, $p = .14$, Hedge's $g = -0.37$; VO: $t(15) = 1.83$, $p = .09$, Hedge's $g = 0.43$; AV: $t(15) = -.59$, $p = .57$, Hedge's $g = -0.14$). To investigate whether PTL Mouth differed between conditions, post-hoc paired-sample *t*-tests with adjusted alpha levels of .05/3 were conducted. These tests showed that, when presented with ADS, adult participants' relative attention to the speaker's mouth region was greater for VO than for AO, and VO than AV conditions (VO vs. AO: $t(15) = 5.29$, $p < .001$, Hedge's $g = 0.82$; VO vs. AV: $t(15) = 4.01$, $p = .001$, Hedge's $g = 0.56$) but the difference between AO and AV conditions was not significant ($t(15) = 2.42$, $p = .029$, Hedge's $g = 0.23$).

When ADS was compared to IDS, post-hoc paired-sample *t*-tests revealed that adult participants spent a significantly greater proportion of time attending to the speaker's mouth region during IDS than during ADS in the AV and VO conditions (AV: $t(15) = 2.83$, $p = .013$, Hedge's $g = 0.32$; VO: $t(15) = 2.63$, $p = .019$, Hedge's $g = 0.40$; AO: $t(15) = 2.17$, $p = .05$, Hedge's $g = 0.33$).

Table 15

Means (and Standard Deviations) of Attention and Preferential Looking to the Speaker's Mouth for ADS and IDS

	ADS		IDS	
	n	**Mean (SD)**	***n***	**Mean (SD)**
AO				
Attention	16	76.73 (12.90)	16	74.93 (12.08)
PTL Mouth	16	38.04 (30.40)	16	53.38 (30.40)
VO				
Attention	16	84.58 (13.24)	16	80.11 (13.89)
PTL Mouth	16	62.41 (27.11)	16	75.52 (27.11)
AV				
Attention	16	82.06 (13.35)	16	80.34 (11.50)
PTL Mouth	16	45.39 (31.47)	16	59.92 (31.47)

5.1.3.3 Attention and Cortical Tracking.

To investigate the relationship between attention and cortical tracking of ADS, Pearson's correlations between attention and stimulus reconstruction accuracy were conducted for each condition. Attention and stimulus reconstruction accuracy were significantly correlated in VO and AV conditions (VO: $r(15) = .52$; $p = .04$; AV: $r(15) = .75$, $p < .001$), but not in AO condition ($r(15) = -.02$; $p = .95$), indicating that greater attention to the screen was associated with more accurate envelope tracking when VO and AV speech were presented.

5.1.3.4 PTL Mouth and Cortical Tracking.

To investigate the relationship between individual differences in looking preference to the speaker's mouth region and cortical tracking of ADS, Spearman's correlations between looking preference and stimulus reconstruction accuracy were conducted for each condition. None of these correlations were significant (AO: $r(15) = -.22$; $p = .42$; VO: $r(15) = -.34$; $p = .20$; AV: $r(15) = -.26$; $p = .33$).

5.1.4 Discussion

Part I of the neurophysiological-behavioural exploration concerned the relationship between looking behaviour to a talking face and cortical tracking of the speech envelope. Several interesting and notable findings emerged.

Overall attention to the screen was generally greater when visual speech information was available. Fixation durations to the screen during presentations of auditory-visual speech were greater than auditory-only (4-year-olds and adults) and visual-only speech (5-month-olds and 4-year-olds). Adults, by comparison, attended similarly to the screen during visual-only and auditory-visual speech. Next, when proportion of time spent looking to the speaker's mouth region relative to the eye region was inspected, differences between conditions were found for 5-month-olds and adults. Even though the proportion of time spent

attending to the speaker's mouth region was not significantly different from chance, 5-month-olds' relative attention to the speaker's mouth region was significantly greater when visual speech information was accessible than when it was not. These findings corroborate past studies that have shown that 6-month-olds fixated more on the speaker's mouth when presented with dynamic videos of the speaker's talking face than when presented with static faces (Shic et al., 2014), and add to the growing body of evidence that infants' gaze behaviour to a talking face reflects their understanding that the mouth conveys important articulatory information (e.g., Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013). Likewise, adults spent a significantly larger proportion of time fixating on the speaker's mouth region when visual speech is available than when it is not. In line with previous findings (e.g., Lansing & McConkie, 2003), the proportion of time that adults spent attending to the speaker's mouth region compared to the eye region was significantly greater than chance only when faced with a silent talking face. While the proportion of time spent fixating on the speaker's mouth was not significantly different between conditions for four-year-olds, the results are limited by the small sample size (n = 8).

Examination of the relationships between gaze behaviour and cortical tracking indicated that only 5-month-olds' relative attention to the speaker's mouth region was positively correlated with the accuracy of cortical tracking of visual-only speech. At first glance, this finding may seem counter-intuitive especially since the relationship between overall attention to screen and cortical tracking accuracy was not significant. However, numerous behavioural studies have shown an attentional shift from the speaker's eyes to the speaker's mouth from around 4 to 8 months of age (e.g., Lewkowicz & Hansen-Tift, 2012; Pons et al., 2015), and that individual variations in attention to a speaker's mouth is related to language acquisition (e.g., Tenenbaum et al., 2015). 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 & Myowa, 2016). Failure to attend to the speaker's mouth is associated with later language learning disorders (Pons et al., 2019). (In this regard, it is surprising that correlations for 4-year-olds or adults were not significant. The former may be due to the small sample size. For the adults, see below regarding IDS vs. ADS). Taken together, attention to a talker's mouth is related to speech processing per se, and overall attention to the screen may not reflect speech processing as accurately as relative attention to the talker's mouth.

The significant relationship between proportion of time spent fixating on the speaker's mouth and 5-month-olds' cortical tracking accuracy of visual-only speech is a critical novel finding. In addition to demonstrating that attention to the speaker's mouth region is directly related to speech processing per se, it bespeaks of the early connection between brain and behaviour—in the absence of an auditory signal, individual differences in infants' looking behaviour to a speaker's talking face influence the accuracy with which neural oscillations synchronise to the speech envelope.

A separate examination of adults' looking behaviour in response to IDS and ADS revealed that different patterns of looking behaviour emerged as a function of speech type. When speech is 'atypical' as in IDS spoken to adults, adults looked longer at the screen in AV and VO conditions compared to AO condition, suggesting that they attended more to the screen when visual information is available. When speech is 'typical' as in ADS to adults, adults attended to the screen more when only visual speech information is available. These differences cannot be explained by attention because attention was similar across conditions for both speech types. Additionally, adults spent a greater proportion of time fixating on the speaker's mouth region relative to the eye region in auditory-visual and visual-only conditions during IDS compared to during ADS, and the proportion of time spent fixating on the speaker's mouth was significantly greater than chance level during IDS when only visual

information was available. These findings are in line with previous research that has established that adult looking behaviour to the speaker's face is context-dependent. Under normal listening conditions, adults attend to the speaker's eyes more than the mouth (Vatikiotis-Bateson et al., 1998). However, in challenging speech processing situations such as when the speech signal is degraded (Vatikiotis-Bateson et al., 1998), or when speech is in a non-native (Barenholtz et al., 2016; Birulés et al., 2020) or an artificial language (Lusk & Mitchel, 2016), adults deploy greater attention to the talker's mouth. Additionally, adults fixate on the mouth region when instructed to identify words (Buchan et al., 2007; Lansing & McConkie, 1999) but direct their gaze to the eyes when tasked with identifying emotions (Buchan et al., 2007) and intonation (Lansing & McConkie, 1999). As adult gaze behaviour to specific facial regions is driven by the parts of the face that provide most information for the task at hand, the observed differences between adults' looking behaviour during IDS and ADS imply that visual speech cues, particularly from the speaker's mouth, were deemed by adults to have facilitatory effects on auditory speech processing of IDS but not ADS.

The investigation of the potential relationships between gaze behaviour and cortical tracking revealed that adults' overall attention was positively correlated with cortical tracking accuracy of visual-only and auditory-visual ADS. This has three key implications.

First, when considered alongside the lack of a significant relationship between relative attention to the speaker's mouth and cortical tracking accuracy of visual-only and auditory-visual speech, this implies that visual information, presumably from broad facial motion occurring over the speaker's face, augments neural speech processing. Alternatively, as postulated by Vatikiotis-Bateson et al. (1998), adults can acquire sufficient information from the talker's mouth through peripheral vision and thus overall attention to the screen may reflect the amount of visual information accumulated from the speaker's mouth movements. The underlying assumption of this claim is that the mouth region is the main driving force

behind the visual speech benefit. If this is indeed the case, then relative attention to the speaker's mouth should also be positively correlated with cortical tracking of visual-only and auditory-visual ADS. Additionally, since adults spent a greater proportion of time attending to the speaker's mouth in visual-only and auditory-visual IDS than ADS, the relationship between relative attention to the speaker's mouth and cortical tracking of visual-only and auditory-visual IDS should be significant. However, the relationships between relative attention to the speaker's mouth and cortical tracking accuracy of visual-only and auditory-visual speech were not significant for either speech type. Therefore, a tentative conclusion is that the available auditory information supplied by the simple ADS stimuli in the current study was sufficient for adults to efficiently decode the speech signal and that any additional visual information that can be acquired from attending specifically to the speaker's mouth was not essential.

Second, the lack of significant correlations between gaze behaviour and cortical tracking of IDS indicate that the facilitatory effects that visual speech cues have on auditory speech processing is likely dependent on the speech type. Infant-directed speech is characterised by exaggerated facial movements and expressions which may be unnatural and unfamiliar to adults. This may interfere with adults' ability to effectively obtain additional acoustic and temporal information from visual speech cues. Non-native speech stimuli that have drastically different temporal properties can be used in the future to investigate this possibility. When coupled with the finding that adults spent a significantly greater proportion of time attending to the speaker's mouth during visual-only and auditory-visual IDS than 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 questionable.

Third, infants and adults likely rely on different visual speech cues to enhance speech perception. The finding that infants' cortical tracking accuracy of visual-only speech was significantly related to their relative attention to the speaker's mouth but not to their overall attention suggests that visual speech cues provided by the talker's mouth movements were particularly useful in assisting infants' neural processing of the silent speech signal. The opposite pattern observed in adults—cortical tracking accuracy of visual-only ADS was significantly related to overall attention but not to relative attention to the speaker's mouth—suggests that visual speech cues provided by the talker's broad facial movements benefited adults' neural processing of the silent speech signal more. As the speech stimuli used in this study were of low complexity, temporal information from the broad movements of the talking face may play a larger role than mouth movements in enhancing adult speech perception. This is suggested by behavioural findings that speech perception is not disrupted even when a spatial low-pass filter is applied to videos of the moving talking face as long as temporal characteristics are kept intact (de Paula et al., 2006; Munhall, Jones, et al., 2004; Munhall, Kroos, & Vatikiotis-Bateson, 2001; Munhall, Kroos, et al., 2004) and that speech perception can be enhanced even with just point-light displays of a talking face (Rosenblum et al., 1996). However, that adults may still be acquiring visual speech information mainly from the talker's mouth movements by way of peripheral vision is a possibility that cannot be entirely ruled out here. As for infants, observing the opening and closing of the mouth may be more beneficial to them because mouth movements provide refined information regarding the start and end points of syllables which can assist in word segmentation since the opening

and closing of the mouth are similar to the syllabic timescale of speech (Chandrasekaran et al., 2009).

Within the limitations of the current investigation, the specific mechanism(s) underlying the observed facilitatory effects that gaze behaviour brings to speech perception remain speculative. To explore further whether directing attention to the mouth augments speech perception, non-native language speech stimuli could be used (e.g., Birulés et al., 2020). If relative attention to the mouth facilitates speech perception, then cortical tracking accuracy to non-native speech will be better in individuals who pay greater relative attention to the speaker's mouth. By contrast, no such facilitatory effects should be observed if this attentional shift is merely a compensatory behaviour that stems from the knowledge that the mouth is the single source of speech. A lack of a significant relationship would also hint at the possibility that adults are not actually registering mouth movement-related information from peripheral vision—they may simply be better off attending to the broad movements occurring over the talker's face. Using both native and non-native speech stimuli in combination with videos depicting only the top half of the speaker's talking face and videos of only a speaker's talking mouth placed in peripheral vision may first clarify whether adults are indeed using peripheral vision to acquire visual information from the mouth and second, reveal the extent to which information from peripheral vision is sufficient for an augmentation of auditory speech perception to occur. The use of non-native speech stimuli in a cortical tracking paradigm would highlight the appeal of neurophysiological measures—no overt behavioural response indicating speech comprehension is necessary. Temporal and spatial properties of visual speech can also be manipulated by means of low-pass filters to investigate their relative importance in the relationship between gaze behaviour and speech perception. A comparison across ages would elucidate any developmental changes in the relative emphasis placed on these temporal and spatial characteristics. Finally, results from

the current investigation could be made more powerful with larger sample sizes, especially for the group of four-year-olds.

To effectively make use of visual speech cues, listeners must know what to attend to. Studies of gaze behaviour have made clear that, from very early on, humans direct their gaze to the speaker's mouth region in part to gather linguistic information (e.g., Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013). The current investigation of gaze behaviour buttresses these studies by showing that relative attention to the speaker's mouth increases when the speaker's talking face is made available to 5-month-olds and adults. Studies on visual speech perception have posited that the redundant temporal properties of visual speech cues are crucial in augmenting speech perception because visual and auditory speech signals share similar temporal modulations (e.g., Munhall, Jones, et al., 2004a; Yehia et al., 1998). To date, investigations of gaze behaviour and auditory-visual speech perception have largely been kept separate (cf. Rennig et al., 2020) even though studies on 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 leaves unanswered many questions regarding the complex web of variables involved in auditory-visual speech perception. In simultaneously recording gaze and EEG data as participants were presented with auditory-only, visual-only, and auditory-visual speech, the present investigation takes a first step toward teasing apart the interactions between a listener's looking behaviour and subsequent speech perception by drawing a direct link between gaze behaviour and the neural processing in auditory-visual speech perception.

5.2 Part II: Cortical Tracking and Later Word Segmentation

The first part of this investigation revealed that looking behaviour to a talking face influences speech perception, suggesting that the type and amount of information extracted from visual speech cues may affect how accurately neural oscillations synchronise with the speech envelope. In the next part of the investigation, the influence of cortical tracking on later behavioural speech segmentation is explored. Specifically, cortical tracking of the speech envelope at 5 months of age is analysed in relation to later word segmentation at 7 months.

Speech is hierarchically organised, with a structure that corresponds remarkably to the organisation of neural oscillations (Ghitza, 2011; Poeppel, 2003). Generally, phonetic features correspond with gamma (>50Hz) and beta (15-30Hz) oscillations, syllables and words with theta (3-8Hz) oscillations, and phrases with delta oscillations (<2Hz). The correspondence between the multiple timescales of speech and the frequency bands of neural oscillations has inspired numerous interrogations of the neural representation of speech. These studies have established a link between cortical tracking of the speech envelope and speech comprehension (e.g., Ding & Simon, 2013; 2014; Doelling et al., 2014; Ghitza, 2012; Ghitza & Greenberg, 2009; Peelle et al., 2012; Zoefel et al., 2018), with more recent studies using electrical brain stimulation to demonstrate the causal role that cortical tracking plays in speech perception—augmenting cortical tracking of the auditory speech envelope improves speech intelligibility (Riecke et al., 2018; Wilsch et al., 2018).

To effectively decode the speech signal, accurate segmentation of the signal must first occur (Doelling et al., 2014; Ghitza, 2017). Oscillation-based models have argued for a cortical computation principle by which speech is decoded in a time-varying window structure that is in synchrony with the multiple time scales of speech (e.g., E. Ahissar & M. Ahissar, 2005; Gross et al., 2013; Lakatos et al., 2005; Poeppel, 2003). According to this

principle, the time-varying window structure is obtained through segmentation which occurs at the syllabic and prosodic (phrasal) levels (Ghitza, 2011; 2017; Poeppel, 2003). In other words, word segmentation performance may be inferred from the accuracy with which neural oscillations synchronise with the speech envelope. In support of this, recent studies have demonstrated the relationship between envelope tracking and concurrent segmentation of an artificial language in infants (D. Choi et al., 2020) and adults (Batterink & Paller, 2017; 2019). Infants and adults who showed stronger envelope tracking during familiarisation to the artificial language were also better at differentiating between words from the artificial language and their non-word foils during the test phase. The relationship between infant cortical tracking and concurrent word segmentation suggests that cortical tracking may be a broad indicator of speech perception capacities that may account for variations observed in other speech perception tasks. For example, it is conceivable that 6-month-old infants who show more accurate cortical tracking of the speech envelope will also show greater recognition of words since the accuracy of cortical tracking is positively related to segmentation performance at 6 months (D. Choi et al., 2020), an age when infants can recognise highly salient words (Bergelson & Swingley, 2012). When considered alongside evidence of the enduring effects that early individual differences have on later language abilities (e.g., Newman et al., 2015), it is not far-fetched to consider that early cortical tracking may be an indicator of general speech perception capacities which translate to later performance differences in speech perception tasks. In Part I of this investigation, 5-month-olds who fixated more on the talker's mouth (vs. eyes) showed better cortical tracking accuracy of silent speech. While the specific mechanism(s) underlying the observed relationship between looking behaviour and cortical tracking cannot be determined here, that finding suggests that the varying levels of speech perception at 5 months may already give some infants an edge in mastering other complex tasks necessary for successful language

acquisition. To address this possibility, an exploratory investigation of the relationship between cortical tracking and later word segmentation was conducted here. A subset of 5-month-old infants who participated in Study 1 also participated in the AV condition in Study 2. It is hypothesised that cortical tracking at 5 months is a predictor of later word segmentation—infants who show more accurate stimulus reconstruction accuracy in the AV condition will also be better at segmenting continuous AV speech. A second hypothesis relates to the use of visual speech information: infants who show greater visual speech benefit, as quantified by the additive model criterion [i.e., AV > (A+V)], are likely better at extracting visual speech information and may better use this information to augment their segmentation of continuous speech.

5.2.1 Methods

5.2.1.1 Participants.

Of the 18 five-month-olds included in Study 1, 12 participated in the auditory-visual condition of Study 2 after turning 7 months of age (mean age = 7.37 months, range = 7.07 to 7.67 months; 5 males).

5.2.2 Results

To examine whether cortical tracking of the speech envelope at 5 months predicts word segmentation performance at 7 months, a simple linear regression analysis was conducted with AV stimulus reconstruction accuracy obtained from Study 1 as the predictor variable and word segmentation performance obtained from Study 2 (overall *d* scores obtained by averaging *d* scores across Blocks 1 and 2) as the outcome variable. The regression model was not significant, adjusted $R^2 = .012$, $F(1,10) = 0.12$, $p = .74$, $\eta_p^2 = .01$, and AV stimulus reconstruction accuracy was not a significant predictor of overall word segmentation performance, $b = 0.61$, $t(10) = 0.35$, $p = .74$.

To examine whether the extent of visual speech benefit reflected in cortical tracking at 5 months predicts word segmentation performance at 7 months, visual speech benefit was first computed by subtracting (A+V) from AV stimulus reconstruction accuracy. Larger values reflect greater visual speech benefit. These values were then entered into a linear regression model as the predictor variable, and with word segmentation performance as the outcome variable. The regression model was not significant, adjusted R^2 = -.022, $F(1,10)$ = 0.76, p = .40, η_p^2 = .07, and visual speech benefit in cortical tracking at 5 months was not a significant predictor of overall word segmentation performance at 7 months, b = -2.25, $t(10)$ = -0.87, p = .40.

5.2.3 Discussion

Part II of the neurophysiological-behavioural exploration of auditory-visual speech perception in infants examined the relationship between cortical tracking of auditory-visual speech at 5 months and subsequent auditory-visual word segmentation performance at 7 months. Surprisingly, this relationship was not significant. Visual speech benefit in cortical tracking was not a significant predictor of later word segmentation performance.

Behavioural evidence suggests that infants at 6 months already show some rudimentary form of speech segmentation (Bergelson & Swingley, 2012), and a very recent neurophysiological study that investigated statistical learning in infant word segmentation (D. Choi et al, 2020) demonstrated the relationship between neural oscillations and concurrent word segmentation in 6-month-olds. In that study, EEG data were simultaneously recorded as infants were familiarised with an artificial language during a behavioural word segmentation task. The researchers found that the time course of cortical tracking in 6-month-olds reflected learning: cortical tracking was stronger at the syllable level during the start of familiarisation but the synchronisation of neural oscillations to word frequency grew stronger over the course of familiarisation, indicating a progressive acquisition of the words contained in the

artificial language. Importantly, the researchers found that 6-month-old infants who showed stronger cortical tracking at the word level during familiarisation to an artificial language also show better segmentation of the same artificial language (D. Choi et al., 2020). Crucially, these findings explicate how individual variations in cortical tracking is reflected at the behavioural level in auditory-only speech perception. With these findings in mind, it is surprising that 5-month-olds' cortical tracking of the speech envelope here was not associated with their later word segmentation performance.

Several limitations preclude a more definite conclusion. First, cortical tracking here was quantified over a frequency range of 0.1 to 12Hz. This range was selected to reflect the frequency bands commonly used in investigating speech perception (e.g., Crosse, Di Liberto, & Lalor., 2016; Jessen et al., 2019; Kalashnikova et al., 2018). As segmentation is postulated to occur at the syllabic and phrasal levels (e.g., Ghitza, 2017), and as learning is reflected in the progressive increment in cortical tracking accuracy at the word level relative to the syllabic level (Buiatti, Peña, & Dehaene-Lambertz, 2009), the relationship between cortical tracking and later segmentation may be made clearer if, as in D. Choi et al. (2020), a *Word Learning Index* (WLI; Batterink & Paller, 2017; 2019)—the ratio of cortical tracking at the word level to cortical tracking at the syllable level—were used. Second, to directly test whether the ability to integrate auditory and visual information is related to segmentation of continuous auditory-visual speech, infants should also participate in an auditory-only word segmentation task. If the difference in segmentation of auditory-only and auditory-visual speech is positively associated with the difference in AV vs. (A+V) stimulus reconstruction accuracy, then it would indicate that individual differences from early on may have a compounding effect on language development. Finally, here the sample size was small (n = 12); increased sample size in future tests would provide more power.

Although the present exploration is limited and did not reveal a significant relationship between cortical tracking and later word segmentation, it is included here as an exploratory endeavour in order to investigate the role that individual differences might play in the interactions between neurophysiological and behavioural indices of auditory-visual speech perception. Many questions remain unresolved. Can attention to visual speech cues serve as a compensatory strategy for infants who have weaker statistical learning capacities? Or, are infants' statistical learning abilities a reflection of a general cognitive capacity that affects their ability to extract information from visual speech cues? If cortical tracking can be artificially augmented (Riecke et al., 2018), can this then aid infants' language development? The recent advancements in technology did not just widen the scope of research, they also foreground the gaps in the literature that have yet to be addressed.

5.3 Summary

A two-pronged approach was taken to explore the potential links between neural and behavioural indices of auditory-visual speech perception. Part I examined whether looking behaviour to a talker's face influences cortical tracking of the speech envelope. Crucially, 5-month-olds' relative attention to a talker's mouth was positively related to cortical tracking of visual-only speech, while adults' overall attention to the screen was positively related to cortical tracking of visual-only and auditory-visual speech. Part II examined whether cortical tracking at 5 months predicts segmentation performance of auditory-visual speech two months later, at 7 months. No significant relationship was found. Coupled with previous research, these investigations—while preliminary—necessitate studying the oft-overlooked individual differences to fully appreciate the nuances underlying auditory-visual speech perception.

CHAPTER 6

General Discussion

Most conversations occur face-to-face, and listeners gather information from the speaker's face to assist with decoding the auditory speech signal (e.g., Sumby & Pollack, 1954). To delve deeper into the benefits that visual speech cues bring to speech perception, the thesis approached this issue from three directions. First, the neural processing underlying visual speech benefit was investigated. Cortical tracking of the speech envelope was compared between auditory-only, visual-only, and auditory-visual presentations of continuous speech across 5-month-olds, 4-year-olds, and adults. Second, the advantages of having access to visual speech information were examined in relation to word segmentation, the perceptual segmentation of individual words from continuous speech. Segmentation performance was compared between two groups of 7.5-month-olds who participated in either an auditory-only or an auditory-visual condition of a behavioural word segmentation task. Third, the possible interactions between neurophysiological and behavioural indices (gaze direction and segmentation ability) of auditory-visual speech perception were explored.

In this final chapter, results from the three investigations (Chapters 3-5) of this thesis are first summarised, and then discussed in concert.

6.1 Summary of Results

In the first investigation, EEG and gaze data were recorded simultaneously as 5-month-olds, 4-year-olds and adults were presented with auditory-only, visual-only, and auditory-visual speech. Cortical tracking of the speech envelope was measured by backward (decoders) and forward (temporal response functions) modelling. Across ages, cortical tracking of the speech envelope was poorest when silent speech was presented. Five-month-olds' cortical tracking was most accurate when both auditory and visual speech information

was available, but cortical tracking by 4-year-olds and adults did not differ between auditory-only and auditory-visual conditions, suggesting that visual speech cues provided substantial benefits only for infant participants. Visual speech benefit, when quantified by the additive model criterion [i.e., AV vs. (A+V)], was found for 5-month-olds and adults at the frontal, occipital and temporal scalp locations. To ensure that exposure to IDS, a speech register directed to young infants, was not a confound in the results obtained from adult participants, EEG and gaze data were also collected from adults presented with ADS stimuli. As for IDS, cortical tracking did not differ between auditory-only and auditory-visual ADS. But unlike IDS, cortical tracking was more accurate for auditory-only than visual-only ADS, and there was no evidence of a visual speech benefit for ADS.

In the second study, 7.5-month-olds participated in a behavioural word segmentation task. As the benefits that can be derived from visual speech cues also depend on whether individuals are attending to the specific regions of a talking face, gaze behaviour was simultaneously collected and analysed. The word segmentation task was presented in either auditory or auditory-visual modality using a central screen that accommodated the use of an eye-tracker. Infants in both conditions were able to segment words from a continuous speech stream, but only the group of infants who were presented with auditory-visual stimuli showed successful word segmentation that persisted into the second test block. Examination of gaze behaviour revealed that differences in attention and mouth versus eye preference only emerged between conditions over time; time course analysis of the proportion of time spent fixating on the talker's mouth revealed that infants in the auditory-visual condition spent a larger proportion of time fixating on the mouth at the beginning of trials than infants in the auditory-only condition in Target and Non-Target trials, but this difference diminished over time.

The final study examined the possible interactions between neurophysiological and behavioural indices of auditory-visual speech perception. Part I explored the relationship between looking behaviour to a speaker's face and cortical tracking of the speech envelope. Two main findings were drawn from this investigation. First, the proportion of time that 5-month-olds spent fixating on the speaker's mouth was related to the accuracy of their neural tracking in the visual-only condition. Second, adults' overall attention was related to cortical tracking accuracy in visual-only and auditory-visual ADS conditions (where visual speech information is available). Part II explored the possible relationship between cortical tracking at 5 months and subsequent performance on segmentation of continuous auditory-visual speech at 7 months, but no evidence for such a relationship was found.

These results identify and highlight two key themes. First, visual speech information facilitates speech perception. Second, looking behaviour modulates auditory-visual speech perception. These two themes form the crux of this discussion, and future endeavours that could resolve the speculations raised in this discussion are also considered in detail.

6.2 Visual Speech Information Facilitates Speech Perception

Visual speech benefit, as quantified by the additive model criterion, i.e., AV > (A+V), was evident here but the results were uneven; visual speech benefit was found for 5-month-olds and adults, but not for 4-year-olds, and visual speech benefit was only detected by forward modelling (in the frontal, occipital and temporal scalp regions) and not by backward modelling. Four issues relating to these findings will be discussed.

The first concerns the finding that visual speech benefit was observed in the frontal, occipital and temporal scalp regions using forward modelling, but not when analyses involved the entire scalp via backward modelling. As neuroimaging studies have identified the superior temporal sulcus (STS) as a candidate for the locus of auditory-visual speech integration in infants (Blasi et al., 2011; Nakato et al., 2009) and adults (e.g., Beauchamp et

al., 2010; Deen et al., 2020; Overath et al., 2015; Riedel et al., 2015), it may be that the visual speech benefit is concentrated in the STS. An alternative explanation is that the magnitude of the visual speech benefit was small and was not strong enough to be detected in backward modelling which involves the whole scalp. However, as backward modelling is supposed to be more sensitive to the multivariate nature of speech (Crosse, Di Liberto, Bednar, & Lalor, 2016), firmer conclusions can only be made after future studies that invoke greater reliance on visual speech cues (e.g., speech-in-noise paradigms) are conducted. Further analyses involving the use of other modelling approaches, such as those based on mutual information models (Nock et al., 2003) and hidden Markov models (Rabiner, 1989), will also provide additional insights that will paint a more coherent picture of the present findings.

Second, the finding that five-month-olds benefit from having access to a speaker's talking face suggests that current oscillation-based models of auditory-visual speech perception (Peelle & Sommers, 2015) may be incomplete. Oscillation-based models posit that the onset of 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 & Obleser, 2012). The greater the amount of predictive information provided by visual speech cues, the higher the degree of 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 & Bernstein, 2019). These oscillation-based models hold that phonetic level knowledge is necessary. However, whether 5-month-olds have acquired the phonemic repertoire required for such predictions is still unclear: phonetic acquisition was previously thought to occur during the second half of the first year (Friederici & Wessels, 1993; Jusczyk et al., 1994;

Polka & Werker, 1994), but recent evidence suggests that infants at 3 months already show native-language phonological knowledge (J. Choi, Broersma, & Cutler, 2017; J. Choi, Cutler, & Broersma, 2017). Furthermore, even if phonological acquisition is already in progress by 5 months, whether the knowledge accrued at 5 months is sufficient for phonetic-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 by 8-month-old infants has been found to be better than by adults at frequency bands that corresponded to the rhythmic and phonemic patterns, whereas cortical tracking at the syllabic level did not differ between age groups (Leong et al., 2017). As such rhythm 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 augments speech perception differently for infants and adults: while the phase-reset may inform predictions at the syllable or prosodic rhyme level for infants, the phase-reset additionally serves to inform predictions at the word level for adults. To verify this, phase-locking activity in the frequency bands related to the syllable, word and prosodic levels could, in future studies, be measured for unimodal and multimodal speech stimuli and compared between age groups. Such studies will assist in elucidating the roles that visual speech cues play and their interactions with age, and in addition, assist in the fine-tuning of oscillation-based models of auditory-visual speech perception.

Third, the fact that visual speech benefit was not evident in 4-year-olds warrants more investigation. When 4-year-olds' looking behaviour and cortical tracking are considered in concert, the absence of a visual speech benefit is surprising because 4-year-olds attended most to the screen during auditory-visual presentations. One explanation is that visual speech cues provide additive benefits to speech perception only in difficult listening conditions. Studies with adults have found that adults rely on and benefit from visual speech cues more in difficult listening conditions such as when faced with a non-native language (Birulés et al., 2020) or in a multi-talker scenario (Buchan et al., 2008). This is further supplemented by the demonstration here that visual speech benefit was only observed in adults' responses to IDS and not to ADS, suggesting that visual speech cues had substantial facilitatory effects for adults only when listening conditions are more challenging as when faced with IDS. As the stimuli used here consisted of short sentences that were recorded in a quiet background, it is possible that 4-year-olds, like adults, were able to decode the auditory speech signal effectively, and did not derive substantial benefits from the addition of visual speech cues.

A second explanation is that 4 years may represent an age at which a developmental shift in auditory-visual speech perception occurs. This is suggested by inconsistencies in results from behavioural studies with 4-year-olds: Jerger et al. (2014) found that 4-year-olds were better at identifying words that were presented in auditory-visual than in auditory-only conditions when the auditory onset of the words were non-intact while Maidment et al. (2015) found that 4-year-olds did not comprehend noise-vocoded sentences better when visual speech cues were added. Although such inconsistencies may stem from task demands because noise-vocoded sentences (vs. non-intact words) may have placed additional demands on 4-year-olds' speech processing capabilities, the absence of a visual speech benefit in 4-year-olds' cortical tracking even when no overt response was required suggests that there is more to it than meets the eye. Given that the auditory-visual speech perception literature has

demonstrated that the visual speech benefit is present in infancy (Hollich et al., 2005; Teinonen et al., 2008), and that this benefit increases with age (adults vs. 6- to 8-year-olds: Lalonde & Holt, 2015; 12- to 14-year-olds vs. 8- to 11-year-olds vs. 5- to 7-year-olds: Ross et al., 2011; 6- to 10-year-olds vs. 5- to 6-year-olds: Fort et al., 2012), 4-year-olds would have been expected to show a visual speech benefit. Further suggesting a developmental shift is the considerable increase in visual speech influence that appears between 6 to 8 years of age (Sekiyama & Burnham, 2008), which suggests that information from the visual modality is given more weight in this period. Collectively, these findings along with the inconsistencies indicating some evidence that 4-year-olds are able to use visual speech information to bolster speech perception point toward a conceivable re-weighting of the relative roles that different visual speech cues play in auditory-visual speech perception that occurs sometime around 4 years of age.

Alternatively, the observation that 4-year-old participants grew increasingly restless during the experimental session raises the possibility that attention to the task may have played a confounding role. The increasing restlessness that 4-year-olds exhibited suggests that the stimuli used in the EEG may not have sufficiently engaged them. If the stimuli were not engaging for 4-year-olds, then they might have been less motivated to understand the speaker and this, in turn, may have modulated speech processing (Pichora-Fuller et al., 2016) and influenced the results. Thus, until attention is ruled out as a confound, no firm conclusion can be drawn.

These possible explanations remain speculative until future studies designed to address them are conducted. To elucidate whether attention confounded the results from 4-year-olds, future studies could employ more engaging stimuli (e.g., longer story-telling video passages). If attention (or the lack of it) is the underlying reason for the missing visual speech benefit here, then 4-year-olds should show a visual speech benefit in response to more

engaging stimuli. Further, if 4-year-olds show a visual speech benefit, then it would explicate the inconsistent behavioural findings and clarify whether a perceptual re-weighting of visual speech cues occurs at this age. To make clear whether visual speech cues provide additive benefits to 4-year-olds only in difficult listening conditions, future studies could implement a speech-in-noise paradigm. Such a future study could be further supplemented by a comprehension task as comprehension level is a good indicator of attention to the task. If there was a significant relationship between cortical tracking accuracy and comprehension performance, then this would suggest a link between neurophysiological and behavioural indices of speech perception. Finally, to chart the developmental trajectory of the visual speech benefit with greater acuity, future studies could compare cortical tracking of 4-year-olds and an older age group which would allow a clearer picture of the developmental trajectory of the visual speech benefit.

Fourth, no relationship was found between 5-month-olds' cortical tracking and 7.5-month-olds' word segmentation performance. A relationship would have been expected because cortical tracking accuracy is an indicator of concurrent word segmentation performance (Batterink & Paller, 2017; 2019; D. Choi et al., 2020; Ghitza, 2011; 2017; Poeppel, 2003), and because early individual differences in word segmentation ability affect later language abilities (e.g., Newman et al., 2015). One possible reason for the present results is that cortical tracking was quantified over a broad frequency range of 0.1 to 12Hz. This frequency range was chosen in accordance with previous studies with infants (Jessen et al., 2019; Kalashnikova et al., 2018) and adults (Crosse, Di Liberto, & Lalor, 2016). As word segmentation is postulated to occur at the syllabic and phrasal levels (e.g., Ghitza, 2017) and as learning is reflected in the progressive increment in cortical tracking accuracy at the word level relative to the syllabic level (Buiatti et al., 2009), the broad frequency range used may have masked effects pertaining to word segmentation.

Further insights may be gleaned from the findings that visual speech cues from a talker's face (1) increase the accuracy with which 5-month-olds' cortical oscillations track the speech envelope and (2) strengthen 7.5-month-olds' word segmentation. Five-month-olds' cortical tracking of the speech envelope was most accurate in auditory-visual presentations, followed by auditory-only then visual-only conditions. This pattern of results cannot be explained by overall attention to the screen because a significant difference in attention was only found between AV and VO conditions (AV > VO), substantiating that visual speech information facilitated speech perception by increasing the accuracy with which neural oscillations synchronise with the speech envelope. The facilitation by visual speech cues have been posited to be driven by the redundant information that the temporal coherence between auditory and visual speech signals provides (Bahrick & Lickliter, 2000). In comparison, the nature of the more robust segmentation for the AV versus the AO group of 7.5-month-olds—a more enduring ability for segmentation over time—suggests that the memory trace for newly learnt words is strengthened by visual speech cues. When considered in parallel, these findings hint at distinct pathways via which visual speech information augments infant speech perception. This is not to say that these pathways are independent. It is possible that they are intertwined: the increased accuracy with which neural oscillations track the auditory-visual speech envelope may later fortify the memory trace for newly learnt words.

In this thesis, only visual cues from the speaker's talking face were examined. However, Hollich et al. (2005) additionally demonstrated that 7.5-month-olds did not segment speech in noise when it was paired with a still photo of the speaker's face, but were able to segment speech in noise when a temporally congruous oscilloscope pattern or the speaker's talking face was paired with the auditory speech recordings. This suggests that the temporal alignment between auditory and visual cues drives the augmentation of word

segmentation in Hollich et al. (2005) and implies that the same may also hold true for cortical tracking accuracy. In other words, temporal coherence between visual cues—that need not necessarily be from the speaker's talking face—and the auditory signal may augment 5-month-olds' neural oscillatory activity to the same extent as visual speech cues. Although this line of reasoning seems tenable, evidence that a speaker's talking face provides speaker-specific indexical cues that support word learning indicate that there may be more to a speaker's talking face than the simple temporal congruence between facial movements and the auditory speech signal. Infants are sensitive to acoustic indexical cues (e.g., Houston & Jusczyk, 2000; Schmale et al., 2010) and can use such cues in later word recognition (van Heugten & Johnson, 2012); and adults make use of visual indexical cues to learn an artificial language (Mitchel & Weiss, 2010). In light of these findings, it is reasonable to postulate that visual speech cues additionally contain speaker-specific indexical properties which may play a subtle but important role in auditory-visual speech perception.

To address these speculations, a future study could incorporate cortical tracking measures into a word segmentation task. Such a study could include an AO condition that pairs the auditory recordings with a still image of the speaker's face, and two AV conditions that pair the auditory recordings with either the corresponding video of the speaker talking or a temporally congruous oscilloscope pattern (Hollich et al., 2005). As 6-month-olds' word learning of an artificial language is reflected by a transition from stronger cortical tracking at the syllable-level frequency band to stronger cortical tracking at the word-level frequency band (D. Choi et al., 2020), visual speech cues may augment word segmentation by accelerating this shift in cortical tracking. Following D. Choi et al. (2020), the transition can be quantified by modelling a learning curve of the *Word Learning Index,* or the ratio of cortical tracking at the word level to cortical tracking at the syllable level, over the familiarisation phase. If visual speech cues facilitate word segmentation by accelerating the

transition of stronger cortical tracking at the syllable level to stronger cortical tracking at the word level, then infants in the AV conditions would show a steeper learning curve compared to infants in the AO condition. To investigate whether visual indexical cues provide additional facilitatory effects, the learning curves of infants from the two AV groups could be compared. If speaker-specific visual indexical cues provide additional benefits to word segmentation, then the gradients of the learning curves of infants who were presented with a speaker's talking face would be steeper than those of infants who were presented with a synchronous oscilloscope pattern. Moreover, because speakers vary in their articulatory movements, seeing a speaker's talking face increases exposure to such idiosyncrasies, and this may then improve the precision with which neural oscillations track the speech envelope. To explore this, cortical tracking accuracy could be compared between the two AV groups. If visual indexical cues serve to finetune cortical tracking accuracy, then cortical tracking accuracy would be more accurate for the group of infants who were presented with the speaker's talking face. If cortical tracking accuracy for the AV groups were found to be greater than the AO group, but not different between the two AV groups, then it would suggest that temporally synchronous cues of a different modality are sufficient for an augmentation of cortical tracking to occur. Such a future study would also expound the acuity of quantifying cortical tracking over a broad frequency range in an examination of word segmentation.

6.3 Looking Behaviour Modulates Auditory-Visual Speech Perception

An analysis of gaze data to the speaker's face that were simultaneously recorded with the EEG data (Study 1) revealed that all age groups attended more to the screen in AV and VO conditions than in the AO condition. It is possible that participants fixate on the screen more in AV and VO conditions because the dynamic movements of the speaker's talking face are visually salient (not only mouth movements, but head and eyebrow movements).

However, analyses pertaining to participants' relative attention to the speaker's mouth region suggest that visual salience is not the *complete* explanation. Infants and adults direct their attention to the mouth when visual speech information is available (AV and VO conditions), indicating that the looking patterns observed here are largely part of an information-seeking strategy. Infants (Tenenbaum et al., 2013), like adults (Lansing & McConkie, 2003), understand that the mouth is the single source of auditory speech information and will redirect their attention to the speaker's mouth to gather linguistic information. This interpretation is further bolstered by the time course plots of 7.5-month-olds' attention to the speaker's mouth during the word segmentation task: fixation durations to the speaker's mouth by infants in the auditory-visual condition increased over time during familiarisation passages but decreased over time in test trials that consisted of repetitions of isolated word tokens. This suggests that infants had acquired sufficient linguistic information and no longer needed to attend specifically to the speaker's mouth. This pattern of looking behaviour is similar to that of adults: Lusk and Mitchel (2016) found that adults' fixation durations to a speaker's mouth decreased as they became more familiar with an artificial language spoken by the talker. Such an information-seeking strategy is in accordance with the cognitive relevance hypothesis (Henderson et al., 2009) which postulates that visual attention is driven by current information-gathering needs more so than by visual salience.

When looking behaviour was compared across the three conditions (AO, VO, and AV), subtle differences emerged. Five-month-old infants fixated on the speaker's mouth more when visual speech information is available (VO and AV conditions). On the other hand, adults fixated more on the speaker's mouth in IDS than ADS presentations, and similar to the infants, when visual speech information is available, but especially when *only* visual speech information is available. These behavioural patterns are in accord 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 because they lack the expertise in their native language in order to rely mainly on the auditory signal (Lewkowicz & Hansen-Tift, 2012), whereas adults' relative attention to the speaker's mouth region increases in challenging speech processing situations (Birulés et al., 2020) when the auditory signal cannot be relied on (as when faced with unfamiliar IDS). So, 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.

If mouth-looking is an information-seeking strategy, then one would expect the relationship between (1) attention to the mouth and adults' cortical tracking accuracy of IDS, and (2) attention to the mouth and infant word segmentation performance to be significant. This was not the case. Interestingly, previous studies with adults have found that adults' looking behaviour is not related to speech recognition performance (Buchan et al., 2007, 2008). This indicates that the present results for adults should not be entirely unexpected. However, this calls into question the utility of such an information-seeking strategy where adults redirect their attention to the speaker's mouth region especially in difficult listening conditions.

In contrast, infants' looking behaviour to a speaker's mouth has been associated with linguistic abilities in past studies (Tenenbaum et al., 2013; 2015; Tsang et al., 2018; Young et al., 2009). One likely explanation for the absence of a relationship between looking times to the speaker's mouth and word segmentation performance is that the stimuli used in the word segmentation study here were words that infants commonly encounter in their everyday lives. Given that infants redirect their attention to the mouth as part of an information-seeking strategy to gather linguistic information, if infants are already familiar with these words, then

there is less motivation for them to attend additionally to the mouth. It is possible then that this may consequently diminish any modulating effect that looking to a speaker's mouth has on word segmentation performance. This explanation remains speculative until a similar word segmentation study that involves novel non-words or an artificial language is conducted.

It is notable that attention to the mouth predicted cortical tracking of visual-only speech at 5 months but did not predict word segmentation performance at 7.5 months. To resolve this seeming discrepancy, a future study is required in which a visual-only condition is implemented in the behavioural word segmentation task conducted in Study 2. Such an implementation would require clever manipulation because a silent video may not hold infants' attention for long. Nevertheless, a visual-only condition could prove informative. If relative attention to the speaker's mouth predicts infants' segmentation performance of silent speech, then it would suggest that, while visual speech cues from the speaker's mouth augment word segmentation, visual speech information may only serve to strengthen already-segmented words when infants have access to a clear auditory signal as in the auditory-visual condition.

More work is necessary to clarify the 4-year-olds' results. Four-year-olds' attention to the speaker's mouth region did not differ across conditions. As linguistic expertise is posited to be responsible for the shift from attending more to the speaker's mouth to attending more to the speaker's eye region at 12 months (Lewkowicz & Hansen-Tift, 2012), it was expected that, similar to adults, 4-year-olds' relative attention to the speaker's mouth region would be greatest when only visual speech information was available. It is possible that 4-year-olds are more attuned to infant-directed than adult-directed speech as the same exaggerated properties can also be found in child-directed speech, albeit in an attenuated manner (Stern et al., 1983). If so, then 4-year-olds may be less inclined to redirect their attentional focus to the speaker's

mouth. However, this is unlikely to be the case because adults redirected their attention to the speaker's mouth when presented with silent videos of ADS. As the sample size of 4-year-olds' gaze data to the speaker's mouth is small (n = 8), any interpretation of their results remains inconclusive until further investigations with larger sample sizes are conducted.

6.4 Implications

As visual speech cues improve cortical tracking of the speech envelope and augment word segmentation, there are implications for those who have difficulty segmenting auditory-only speech. Two obvious populations are individuals with hearing loss (HL) and individuals from multilingual backgrounds. Individuals with HL receive impoverished auditory input and may thus learn to use other forms of speech information available in their environment. As has been shown in studies of visual benefit in auditory noise, one way to compensate for a degraded auditory signal is through the visual modality. Accordingly, children and adults with HL show greater reliance and better processing of visual speech information (Bernstein et al., 2001; Rouger et al., 2007; Taitelbaum-Swead & Fostick, 2017).

Multilinguals include individuals who are exposed to more than one language and include those who grew up in a multilanguage environment (e.g., bilinguals) and those who are learning a second language. These individuals rely more on visual speech cues (bilinguals: Navarra & Soto-Faraco, 2007; Wang et al., 2009; L2 learners: Birulés et al., 2020) regardless of their language proficiency (bilinguals: Marian et al., 2018; L2 learners: Birulés et al., 2020). The increased reliance on visual speech cues, particularly from the speaker's mouth, is evident even in infants developing in a bilingual language environment (Ayneto & Sebastián-Gallés, 2017; Fort et al., 2017; Pons et al., 2015). This is argued to be an adaptive mechanism that young bilingual learners develop to disambiguate speech sounds from both languages as fixations on the speaker's mouth are also influenced by how rhythmically- and phonologically-related the two languages are (Birulés et al., 2018).

Both populations highlight that visual speech cues provide an additional source of information that may assist in word segmentation. Even though the reliance on visual speech cues is greater in these populations, and that these populations benefit from the addition of visual speech information (individuals with HL: Grant et al., 1998; Moradi et al., 2016; Rouger et al., 2007; bilingual individuals: Navarra & Soto-Faraco, 2007), it remains unclear whether the visual speech benefit would be greater for these population. Indirect evidence comes from the finding that the visual speech benefit increases as listening conditions become more challenging (Sumby & Pollack, 1954), suggesting that populations who are confronted with challenging listening situations would likely show a larger visual speech benefit. In this regard, it would be interesting to examine whether such gains are compensatory in that they bolster auditory speech perception to reach similar levels as monolingual individuals with normal hearing. Once these possibilities are addressed in future word segmentation studies, and if the visual speech benefit is found to be compensatory, then there would be a strong case for incorporating visual speech cues in language-learning contexts for both HL and multilingual language learners because word segmentation is a basic skill that underlies many aspects of language acquisition.

6.5 Conclusion

In a letter to his friend in May 1783, Benjamin Franklin wrote about his latest invention, the bifocal spectacles:

> "… and when one's ears are not well accustomed to the sounds of a language, a sight of the movements in the features of him that speaks helps to explain, so that I understand French better by the help of my spectacles." (Franklin, 1835, p. 157).

This observation underscores the importance of visual speech information. Concurrent visual speech can augment auditory speech perception in optimal and suboptimal listening conditions. Even so, research in visual speech perception has been overshadowed by

research in auditory speech perception because a greater range of information is generally available through heard than seen speech. To assist in the current movement to redress this imbalance, this thesis examined how visual speech cues augment auditory speech perception over development. The first study illustrated a visual speech benefit in 5-month-olds' and adults' cortical tracking of the speech envelope. The second study showed that while there is successful word segmentation in both auditory and auditory-visual conditions, visual speech information serves to strengthen the memory trace of newly learnt words. The final study showed that idiosyncratic differences in gaze direction have consequences for cortical tracking of the speech envelope and highlighted the importance of accounting for these differences. Together, findings from this thesis articulate the labyrinthine nature of auditory-visual speech perception and expose many gaps in the literature that await future exploration. Critically, the three investigations illuminate the facilitatory effects of visual speech information—*seeing a talking face matters*.

References

Ahissar, E., & Ahissar, M. (2005). Processing of the temporal envelope of speech. In R. Konig, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 295–313). Lawrence Erlbaum Associates, Inc.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372. http://doi.org/10.1073/pnas.201400998

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990. http://doi.org/10.1111/desc.12071

Arizpe, J., Kravitz, D. J., Yovel, G., & Baker, C. I. (2012). Start position strongly influences fixation patterns during face processing: Difficulties with eye movements as a measure of information use. *PloS One*, *7*(2), e31106–17. http://doi.org/10.1371/journal.pone.0031106

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. http://doi.org/10.1038/nn.2810

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Lawrence Erlbaum Associates, Inc.

Ayneto, A., & Sebastián-Gallés, N. (2017). The influence of bilingualism on the preference for the mouth region of dynamic faces. *Developmental Science*, *20*(1), 180. http://doi.org/10.1111/desc.12446

Baart, M., & Samuel, A. G. (2015). Early processing of auditory lexical predictions revealed by ERPs. *Neuroscience Letters*, *585*, 98–102. http://doi.org/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. http://doi.org/10.1016/j.cognition.2013.09.006

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*(2), 190–201. http://doi.org/10.1037//0012-1649.26.2.190

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105. http://doi.org/10.1016/j.cognition.2015.11.013

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. http://doi.org/10.1016/j.cortex.2017.02.004

Batterink, L. J., & Paller, K. A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex*, *115*, 56–71. http://doi.org/10.1016/j.cortex.2019.01.013

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus Is a cortical locus of the McGurk Effect. *Journal of Neuroscience*, *30*(7), 2414–2417. http://doi.org/10.1523/JNEUROSCI.4865-09.2010

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1198. http://doi.org/10.2307/2674075

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921. http://doi.org/10.1073/pnas.1712966114

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. http://doi.org/10.1073/pnas.1113380109

Bergelson, E., & Swingley, D. (2017). Young infants' word comprehension given an unfamiliar talker or altered pronunciations. *Child Development*, *89*(5), 1567–1576. http://doi.org/10.1111/cdev.12888

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*, 386. http://doi.org/10.3389/fnins.2014.00386

Bernstein, L. E., Auer, E. T., & Tucker, P. E. (2001). Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language, and Hearing Research*, *44*(1), 5–18. http://doi.org/10.1044/1092-4388(2001/001)

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2018). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental Science*, *22*(3), e12446. http://doi.org/10.1111/desc.12755

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. *Language, Cognition and Neuroscience*, *92*(2), 1–12. http://doi.org/10.1080/23273798.2020.1762905

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual speech benefit in clear and degraded speech depends on the auditory intelligibility of

the talker and the number of background talkers. *Trends in Hearing*, *23*(5), 233121651983786–14. http://doi.org/10.1177/2331216519837866

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PloS One*, *3*(8), e3022–8. http://doi.org/10.1371/journal.pone.0003022

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G. J., Renvall, V., Deoni, S., Gasston, D., Williams S. C. R., Johnson, M. H., Simmons, A., & Murphy, D. G. M. (2011). Early specialization for voice and emotion processing in the infant brain. *Current Biology*, *21*(14), 1220–1224. http://doi.org/10.1016/j.cub.2011.06.009

Bortfeld, H., Morgan, J. L., & Golinkoff, R. M. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(5), 298–304. http://doi.org/10.1111/j.0956-7976.2005.01531.x

Bosch, L., Figueras, M., Teixidó, M., & Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: Evidence from infants acquiring syllable-timed languages. *Frontiers in Psychology*, *4*, 106. http://doi.org/10.3389/fpsyg.2013.00106

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219. http://doi.org/10.3758/bf03206883

Bristow, D., Dehaene-Lambertz, G., & Mattout, J. (2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. http://doi.org/10.1162/jocn.2009.21076

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13. http://doi.org/10.1080/17470910601043644

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 Research*, *1242*, 162–171. http://doi.org/10.1016/j.brainres.2008.06.083

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage*, *44*(2), 509–519. http://doi.org/10.1016/j.neuroimage.2008.09.015

Burnham, D. (1993). Visual recognition of mother by young infants: Facilitation by speech. *Perception*, *22*(10), 1133–1153. http://doi.org/10.1068/p221133

Burnham, D. (1998). Language specificity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech* (pp. 27–60). Psychology Press.

Burnham, D., & Dodd, B. (1999). Familiarity and novelty preferences in infants' auditory-visual speech perception: Problems, factors, and a solution. In C. Rovee-Collier, L. Lipsitt, & H. Hayne (Ed.), *Advances in Infancy Research* (pp. 170–187). Ablex Publishing Corporation.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. http://doi.org/10.1002/dev.20032

Butler, J., & Frota, S. (2018). Emerging word segmentation abilities in European Portuguese-learning infants: New evidence for the rhythmic unit and the edge factor. *Journal of Child Language*, *45*(6), 1294–1308. http://doi.org/10.1017/S0305000918000181

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. http://doi.org/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 Transactions on Biomedical Engineering*, *67*(4), 1114-1121. http://doi.org/10.1109/TBME.2019.2930186

Choi, D., Batterink, L. J., Black, A. K., Paller, K. A., & Werker, J. F. (2020). Prelingual infants discover statistical word patterns at similar rates as adults: Evidence from neural entrainment. *Psychological Science*, *31*(9), 1161–1173.

Choi, J., Broersma, M., & Cutler, A. (2017). Early phonology revealed by international adoptees' birth language retention. *Proceedings of the National Academy of Sciences, 114*(28), 7307-7312. http://doi.org/10.1073/pnas.1706405114

Choi, J., Cutler, A., & Broersma, M. (2017). Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory. *Royal Society Open Science*, *4*(1), 160660–14. http://doi.org/10.1098/rsos.160660

Chong, S., Werker, J. F., & Russell, J. A. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development*, *12*(3), 211–232. http://doi.org/10.1002/icd.286

Cole, R. A., Jakimik, J., & Cooper, W. E. (1980). Segmenting speech into words. *The Journal of the Acoustical Society of America*, *67*(4), 1323–1332. http://doi.org/10.1121/1.384185

Corrigall, K. A., & Trainor, L. J. (2014). Enculturation to musical pitch structure in young children: Evidence from behavioral and electrophysiological methods. *Developmental Science*, *17*(1), 142–158. http://doi.org/10.1111/desc.12100

Crosse, M. J., & Lalor, E. C. (2014). The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech. *Journal of Neurophysiology*, *111*(7), 1400–1408. http://doi.org/10.1152/jn.00690.2013

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. http://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, *36*(38), 9888–9895. http://doi.org/10.1523/JNEUROSCI.1396-16.2016

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience, 10*, 604. http://doi.org/10.3389/fnhum.2016.00604

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, *96*(3), 233–262. http://doi.org/10.1016/j.cognition.2004.08.005

Danielson, D. K., Bruderer, A. G., Kandhadai, P., Vatikiotis-Bateson, E., & Werker, J. F. (2017). The organization and reorganization of audiovisual speech perception in the first year of life. *Cognitive Development*, *42*, 37–48. http://doi.org/10.1016/j.cogdev.2017.02.004

de Paula, H., Yehia, H. C., Shiller, D., Jozan, G., Munhall, K. G., & Vatikiotis-Bateson, E. (2006). Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visible speech information. In J. Harrington & M. Tabain (Eds.), *Speech production models, phonetic processes, and techniques* (pp. 135–147). Psychology Press.

deBoer, T., Scott, L. S., & Nelson, C. A. (2007). Methods for acquiring and analyzing infant event-related potentials. In M. de Haan (Ed.), *Studies in developmental psychology. Infant EEG and event-related potentials* (pp. 5–37). Psychology Press.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176. http://doi.org/ 10.1126/science.7375928

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behaviour and Development*, *9*, 133–150. https://doi.org/10.1016/0163-6383(86)90025-1

Deen, B., Saxe, R., & Kanwisher, N. (2020). Processing communicative facial and vocal cues in the superior temporal sulcus. *NeuroImage*, *221*, 1–16. http://doi.org/10.1016/j.neuroimage.2020.117191

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. http://doi.org/10.1016/j.jneumeth.2003.10.009

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology*, *66*(1), 85–110. http://doi.org/10.1006/jecp.1997.2379

Desjardins, R., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187–203. http://doi.org/10.1002/dev.20033

Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hearing Research*, *348*, 70–77. http://doi.org/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. http://doi.org/10.1016/j.neuroimage.2018.03.072

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. http://doi.org/10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. http://doi.org/10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735. http://doi.org/10.1523/JNEUROSCI.5297-12.2013

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311. http://doi.org/10.3389/fnhum.2014.00311

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. http://doi.org/10.1038/nn.4186

Dink, J., & Ferguson, B. (2018). eyetrackingR [Computer software manual]. http://www.eyetracking-R.com (R package version 0.1.8.)

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. http://doi.org/10.1016/j.neuroimage.2013.06.035

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.

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass*, *4*(7), 473–480. http://doi.org/10.1111/j.1749-818X.2010.00210.x

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. http://doi.org/10.1126/science.171.3968.303

Elsabbagh, M., Bedford, R., Senju, A., Charman, T., Pickles, A., Johnson, M. H., The BASIS Team. (2013). What you see is what you get: Contextual modulation of face scanning in typical and atypical development. *Social Cognitive and Affective Neuroscience*, *9*(4), 538–543. http://doi.org/10.1093/scan/nst012

Erdener, D., & Burnham, D. (2013). The relationship between auditory-visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology*, *116*(2), 120–138. http://doi.org/10.1016/j.jecp.2013.03.003

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5). http://doi.org/10.2307/1166093

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage*, *186*, 33–42. http://doi.org/10.1016/j.neuroimage.2018.10.057

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. *Journal of Cognitive Neuroscience*, *27*(5), 1060–1067. http://doi.org/10.1162/jocn_a_00764

Fort, M., Ayneto-Gimeno, A., Escrichs, A., & Sebastián-Gallés, N. (2017). Impact of bilingualism on infants' ability to learn from talking and nontalking faces. *Language Learning*, *68*(5), 31–57. http://doi.org/10.1111/lang.12273

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1207–1223. http://doi.org/10.1080/01690965.2012.701758

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *International Journal of Behavioral Development*, *36*(6), 457–467. http://doi.org/10.1177/0165025412447752

Franklin, B. (1835). *The works of Dr. Benjamin Franklin.* Campe.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, *54*, 287–295. http://doi.org/10.3758/bf03205263

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 130. http://doi.org/10.3389/fpsyg.2011.00130

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, *3*, 238. http://doi.org/10.3389/fpsyg.2012.00238

Ghitza, O. (2017). Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*, *32*(5), 545–561. http://doi.org/10.1080/23273798.2016.1232419

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126. http://doi.org/10.1159/000208934

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*, 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy, 2*(2), 219-231. http://doi.org/10.1207/S15327078IN0202_7

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word-object relations. *Infancy*, *9*(3), 259–288. https://doi.org/10.1207/s15327078in0903_1

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. http://doi.org/10.1111/j.1467-9280.2008.02117.x

Goldstein, M. H., Schwade, J., Briesch, J., & Syal, S. (2010). Learning while babbling: Prelinguistic object-directed vocalizations indicate a readiness to learn. *Infancy*, *15*(4), 362–391. http://doi.org/10.1111/j.1532-7078.2009.00020.x

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." *Journal of Neuroscience*, *33*(4), 1417–1426. http://doi.org/10.1523/jneurosci.3675-12.2013

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. http://doi.org/10.1111/infa.12006

Grant, K. W., & Bernstein, J. G. W. (2019). Toward a model of auditory-visual speech intelligibility. In A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Multisensory Processes*. Springer International Publishing.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3 Pt 1), 1197–1208. https://doi.org/10.1121/1.1288668

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677–2690. http://doi.org/10.1121/1.422788

Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1529–1542. http://doi.org/10.1044/1092-4388(2010/09-0005)

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752. http://doi.org/10.1371/journal.pbio.1001752

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of

talking faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1275–1281. http://doi.org/10.1037/xhp0000208

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception & Psychophysics*, *77*(4), 1333–1341. http://doi.org/10.3758/s13414-014-0821-1

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. http://doi.org/10.1016/j.neuroimage.2013.10.067

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. http://doi.org/10.3758/PBR.16.5.850

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences*, *109*(49), 20095–20100. http://doi.org/10.1073/pnas.1213390109

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, *109*(2), 748–763. http://doi.org/10.1121/1.1337959

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, *76*(3), 598–613. http://doi.org/10.1111/j.1467-8624.2005.00866.x

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582. http://doi.org/10.1037//0096-1523.26.5.1570

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review*, *7*(3), 504–509. http://doi.org/10.3758/BF03214363

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13*(4), 341–348. http://doi.org/10.1002/icd.364

Höhle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, *6*(2), 122–127. http://doi.org/10.1111/1467-7687.00261

Hyde, D. C., Jones, B. L., & Flom, R. (2011). Neural signatures of face–voice synchrony in 5-month-old human infants. *Developmental Psychobiology*, *53*(4), 359–370. http://doi.org/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. http://doi.org/ 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. *Developmental Science*, *55*(6), e12825. http://doi.org/10.1111/desc.12825

Irwin, J., Avery, T., Brancazio, L., Turcios, J., Ryherd, K., & Landi, N. (2017). Electrophysiological indices of audiovisual speech perception: Beyond the McGurk effect and speech in noise. *Multisensory Research*, *31*, 39-56. http://doi.org/10.1163/22134808-00002580

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. *International Journal of Pediatric Otorhinolaryngology*, *94*, 127–137. http://doi.org/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. *Journal of Child Language*, *22*, 1–23. http://doi.org/10.1017/s0305000917000265

Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology*, *126*, 295–312. http://doi.org/10.1016/j.jecp.2014.05.003

Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. (2017). Children perceive speech onsets by ear and eye. *Journal of Child Language*, *44*(1), 185–215. http://doi.org/10.1017/S030500091500077X

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. http://doi.org/10.1016/j.neuroimage.2019.116060

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567. http://doi.org/10.1006/jmla.2000.2755

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. http://doi.org/10.1006/cogp.1995.1010

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, *61*(8), 1465–1476. http://doi.org/10.3758/BF03213111

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*(3-4), 159–207. http://doi.org/10.1006/cogp.1999.0716

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants′ sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645. https://doi.org/10.1006/jmla.1994.1030

Kaganovich, N., & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. *Brain and Language*, *139*, 36–48. http://doi.org/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. *Scientific Reports*, *8*(1), 1–8. http://doi.org/10.1038/s41598-018-32150-6

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PloS One*, *9*(10), e111467. http://doi.org/10.1371/journal.pone.0111467

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 Behaviour and Development*, *24*(4), 372–392. http://doi.org/10.1016/S0163-6383(02)00086-3

Knowland, V., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science*, *17*(1), 110–124. http://doi.org/10.1111/desc.12098

Kooijman, V., Hagoort, P., & Cutler, A. (2009). Prosodic structure in early word segmentation: ERP evidence from Dutch ten-month-olds. *Infancy, 14*(6), 591-612. http://doi.org/10.1080/15250000903263957

Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., & Cutler, A. (2013). Predictive brain signals of linguistic development. *Frontiers in Psychology*, *4*, 25. http://doi.org/10.3389/fpsyg.2013.00025

Kothe, C. A. E., & Jung, T. P. (2014). *U.S. Patent Application No. 14/895,440*.

Kubicek, C., de Boisferon, A. H., Dupierrix, E., Loevenbruck, H., Gervain, J., & Schwarzer, G. (2013). Face-scanning behavior to silently-talking faces in 12-month-old infants: The impact of pre-exposed auditory speech. *International Journal of Behavioral Development*, *37*(2), 106–110. http://doi.org/10.1177/0165025412473016

Kubicek, C., Gervain, J., de Boisferon, A. H., Pascalis, O., Loevenbruck, H., & Schwarzer, G. (2014). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behavior and Development*, *37*(4), 644–651. http://doi.org/10.1016/j.infbeh.2014.08.010

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141. http://doi.org/10.1126/science.7146899

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442–11445. http://doi.org/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. *The European Journal of Neuroscience*, *38*(9), 3363–3369. http://doi.org/10.1111/ejn.12317

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the

auditory cortex. *Journal of Neurophysiology*, *94*(3), 1904–1911. http://doi.org/10.1152/jn.00263.2005

Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually salient speech cues. *Journal of Speech, Language, and Hearing Research*, *58*(1), 135–150. http://doi.org/10.1044/2014_JSLHR-H-13-0343

Lalonde, K., & Holt, R. F. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *The Journal of the Acoustical Society of America*, *139*(4), 1713–1723. http://doi.org/10.1121/1.4945590

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech*, *Language, and Hearing Research, 42*(3), 526–539. http://doi.org/10.1044/jslhr.4203.526

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552. http://doi.org/10.3758/bf03194581

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. http://doi.org/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. *Developmental Psychology*, *39*(5), 795–804. http://doi.org/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. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436. http://doi.org/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. http://doi.org/10.1111/infa.12167

Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Frontiers in Psychology*, *7*, 52. http://doi.org/10.3389/fpsyg.2016.00052

Mackensen, N. A., & Grossmann, T. (2015). Learning to match auditory and visual speech cues: Social influences on acquisition of phonological categories. *Child Development*, *86*(2), 362–378. http://doi.org/10.1111/cdev.12320

Maidment, D. W., Kang, H. J., Stewart, H. J., & Amitay, S. (2015). Audiovisual integration in children listening to spectrally degraded speech. *Journal of Speech, Language, and Hearing Research*, *58*(1), 61–68. http://doi.org/10.1044/2014_JSLHR-S-14-0044

Marian, V., Hayakawa, S., Lam, T., & Schroeder, S. (2018). Language experience changes audiovisual perception. *Brain Sciences*, *8*(5), 85–14. http://doi.org/10.3390/brainsci8050085

Marquis, A., & Shi, R. (2008). Segmentation of verb forms in preverbal infants. *The Journal of the Acoustical Society of America*, *123*(4), 203-208. http://doi.org/10.1121/1.2884082

MathWorks (2019). MATLAB: R2019a. *Mathworks, Inc, Natick.*

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465–494. http://doi.org/10.1006/cogp.1999.0721

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

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. *Journal of Vision*, *14*(7), 6. http://doi.org/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. *Journal of Neuroscience*, *35*(22), 8546–8557. http://doi.org/10.1523/JNEUROSCI.4527-14.2015

Mersad, K., & Nazzi, T. (2012). When Mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*(3), 303–315. http://doi.org/10.1080/15475441.2011.609106

Minagawa, Y., Hakuno, Y., Kobayashi, A., Naoi, N., & Kojima, S. (2017). Infant word segmentation recruits the cerebral network of phonological short-term memory. *Brain and Language*, *170*, 39–49. http://doi.org/10.1016/j.bandl.2017.03.005

Mitchel, A. D., & Weiss, D. J. (2010). What's in a face? Visual contributions to speech segmentation. *Language and Cognitive Processes*, *25*(4), 456–482. http://doi.org/10.1080/01690960903209888

Mitchel, A. D., & Weiss, D. J. (2014). Visual speech segmentation: Using facial cues to locate word boundaries in continuous speech. *Language, Cognition and Neuroscience*, *29*(7), 771–780. http://doi.org/10.1080/01690965.2013.791703

Moradi, S., Lidestam, B., & Rönnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy. *Frontiers in Psychology*, *4*, 359. http://doi.org/10.3389/fpsyg.2013.00359

Moradi, S., Lidestam, B., & Rönnberg, J. (2016). Comparison of gated audiovisual speech identification in elderly hearing aid users and elderly normal-hearing individuals:

Effects of adding visual cues to auditory speech stimuli. *Trends in Hearing*, *20*(7), 1–15. http://doi.org/10.1177/2331216516653355

Mullennix, J. W., Bihon, T., Bricklemyer, J., Gaston, J., & Keener, J. M. (2002). Effects of variation in emotional tone of voice on speech perception. *Language and Speech, 45*(3), 255–283. http://doi.org/10.1177/00238309020450030301

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137. http://doi.org/ 10.1111/j.0963-7214.2004.01502010.x

Munhall, K. G., Kroos, C., & Vatikiotis-Bateson, E. (2001). Bandpass filtered faces and audiovisual speech perception. *The Journal of the Acoustical Society of America*, *109*(5), 2314. http://doi.org/10.1121/1.4744133

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics*, *66*(4), 574–583. http://doi.org/ 10.3758/bf03194902

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP Analyses: A step-by-step tutorial review. *Brain Topography*, *20*(4), 249–264. http://doi.org/10.1007/s10548-008-0054-5

Nakato, E., Otsuka, Y., Kanazawa, S., Yamaguchi, M. K., Watanabe, S., & Kakigi, R. (2009). When do infants differentiate profile face from frontal face? A near-infrared spectroscopic study. *Human Brain Mapping*, *30*(2), 462–472. http://doi.org/10.1002/hbm.20516

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. http://doi.org/10.1007/s00426-005-0031-5

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., & Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech, 48*(3), 1–20. http://doi.org/10.1177/00238309050480030201

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19. http://doi.org/10.1006/jmla.2000.2698

Nazzi, T., Mersad, K., Sundara, M., Iakimova, G., & Polka, L. (2014). Early word segmentation in infants acquiring Parisian French: Task-dependent and dialect-specific aspects. *Journal of Child Language*, *41*(3), 600–633. http://doi.org/10.1017/S0305000913000111

Newman, R. S., & Jusczyk, P. W. (1996). The cocktail party effect in infants. *Perception & Psychophysics*, *58*(8), 1145–1156. http://doi.org/10.3758/bf03207548

Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2015). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. http://doi.org/10.1017/S0305000915000446

Newman, R., Bernstein Ratner, N., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychobiology*, *42*(4), 643–655. http://doi.org/10.1037/0012-1649.42.4.643

Nishibayashi, L., Goyet, L., & Nazzi, T. (2014). Early speech segmentation in French-learning infants: Monosyllabic words versus embedded syllables. *Language and Speech*, *58*(3), 334–350. http://doi.org/10.1177/0023830914551375

Nock, H. J., Iyengar, G., & Neti, C. (2003). Speaker localisation using audio-visual synchrony: An empirical study. *International conference on image and video retrieval* (pp. 488–499). Berlin, Heidelberg: Springer.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences, 23*(11), 913–926. http://doi.org/10.1016/j.tics.2019.08.004

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*(1), 156869–9. http://doi.org/10.1155/2011/156869

Ota, M., & Skarabela, B. (2018). Reduplication facilitates early word segmentation. *Journal of Child Language*, *45*(1), 204–218. http://doi.org/10.1017/S0305000916000660

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911. http://doi.org/10.1038/nn.4021

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech*, *Language, and Hearing Research, 28*(3), 381-393. http://doi.org/10.1044/jshr.2803.381

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. *European Journal of Neuroscience*, *50*(8), 3282-3295. http://doi.org/10.1111/ejn.14425

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., Lalor, E. C. (2014). Attentional selection in a

cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. http://doi.org/10.1093/cercor/bht355

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*(33), 255–287. https://doi.org/10.1037/h0084295

Parise, E., & Csibra, G. (2013). Neural responses to multimodal ostensive signals in 5-month-old infants. *PloS One*, *8*(8), e72360. http://doi.org/10.1371/journal.pone.0072360

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*, 14521. http://doi.org/10.7554/eLife.14521.001

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. http://doi.org/10.1111/1467-7687.00271

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. http://doi.org/10.1016/j.cortex.2015.03.006

Peelle, J. E., Gross, J., & Davis, M. H. (2012). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. http://doi.org/10.1093/cercor/bhs118

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. *Scientific Reports*, *6*(1), 34273. http://doi.org/10.1038/srep34273

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314–E3323. http://doi.org/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 & Hearing*, *37*, 5–27. http://doi.org/10.1097/AUD.0000000000000312

Pisoni, D. B. (1981). Some current theoretical issues in speech perception. *Cognition*, *10*(1-3), 249–259. https://doi.org/10.1016/0010-0277(81)90054-8

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, *41*(1), 245–255. http://doi.org/10.1016/S0167-6393(02)00107-3

Polka, L., & Sundara, M. (2012). Word segmentation in monolingual infants acquiring Canadian English and Canadian French: Native language, cross-dialect, and cross-language comparisons. *Infancy*, *17*(2), 198–232. http://doi.org/10.1111/j.1532-7078.2011.00075.x

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421–435. https://doi.org/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. *Psychological Science*, *26*(4), 490–498. http://doi.org/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 Behavior and Development*, *54*, 80–84. http://doi.org/10.1016/j.infbeh.2018.12.003

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences*, *106*(26), 10598–10602. http://doi.org/10.1073/pnas.0904134106

Power, A. J., Colling, L. J., Mead, N., Barnes, L., & Goswami, U. (2016). Neural encoding of the speech envelope by children with developmental dyslexia. Brain and Language, 160, 1-10. http://doi.org/10.1016/j.bandl.2016.06.006

R Core Team. (2020). R: A language and environment for statistical computing. [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/.

Rabiner, L. R. (1989). A Tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review*, *27*(1), 70–77. http://doi.org/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. *Developmental Psychobiology*, *56*(3), 355–372. http://doi.org/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. http://doi.org/10.1371/journal.pone.0169325

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology*, *28*(2), 161–169. http://doi.org/10.1016/j.cub.2017.11.033

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S. J., & Kriegstein, von, K. (2015). Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex, 68,* 1–14. http://doi.org/10.1016/j.cortex.2014.11.016

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1159–1170. http://doi.org/10.1044/jshr.3906.1159

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347–357. http://doi.org/10.3758/bf03211902

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. *European Journal of Neuroscience*, *33*(12), 2329–2337. http://doi.org/10.1111/j.1460-9568.2011.07685.x

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences*, *104*(17), 7295–7300. http://doi.org/10.1073/pnas.0609419104

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. *Human Factors*, *45*(2), 329–336. http://doi.org/10.1518/hfes.45.2.329.27237

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. http://doi.org/10.1126/science.274.5294.1926

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science*, *12*(4), 583–601. http://doi.org/10.1111/j.1467-7687.2009.00809.x

Schmale, R., Cristia, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy*, *15*(6), 650–662. http://doi.org/10.1111/j.1532-7078.2010.00032.x

Schroeder, C. E., Lakatos, P., Kajikawa, Y., & Partan, S. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113. http://doi.org/10.1016/j.tics.2008.01.002

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. http://doi.org/10.1016/j.cognition.2004.01.006

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565–573. http://doi.org/10.1111/j.1467-7687.2006.00534.x

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320. http://doi.org/10.1111/j.1467-7687.2008.00677.x

Shepard, K. G., Spence, M. J., & Sasson, N. J. (2012). Distinct facial characteristics differentiate communicative intent of infant-directed speech. *Infant and Child Development*, *21*(6), 555–578. http://doi.org/10.1002/icd.1757

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, *75*(3), 231–237. http://doi.org/10.1016/j.biopsych.2013.07.009

Simonetti, S., Kim, J., & Davis, C. (2016). Identifying visual prosody: Where do people look?. *Speech Prosody 2016 Conference* (pp.840–844). Boston, MA, USA: Boston University. http://doi.org/10.21437/speechprosody.2016-172

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

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: infant age and context related changes. *Journal of Child Language*, *10*, 1–15. https://doi.org/10.1017/S0305000900005092

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. http://doi.org/10.1121/1.1907309

Taitelbaum-Swead, R., & Fostick, L. (2016). Auditory and visual information in speech perception: A developmental perspective. *Clinical Linguistics & Phonetics*, *30*(7), 531–545. http://doi.org/10.3109/02699206.2016.1151938

Taitelbaum-Swead, R., & Fostick, L. (2017). Audio-visual speech perception in noise: Implanted children and young adults versus normal hearing peers. *International Journal of Pediatric Otorhinolaryngology*, *92*, 146–150. http://doi.org/10.1016/j.ijporl.2016.11.022

Taylor, G., & Herbert, J. S. (2012). Eye tracking infants: Investigating the role of attention during learning on recognition memory. *Scandinavian Journal of Psychology*, *54*(1), 14–19. http://doi.org/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. http://doi.org/10.1016/j.cognition.2008.05.009

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. http://doi.org/10.1111/j.1532-7078.2012.00135.x

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Shah, R. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language

development. *Journal of Child Language*, *42*(6), 1173–1190. http://doi.org/10.1017/S0305000914000725

Thiessen, E. D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, *34*(6), 1093–1106. http://doi.org/10.1111/j.1551-6709.2010.01118.x

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. http://doi.org/10.1037/0012-1649.39.4.706

Thiessen, E. D., & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception & Psychophysics*, *66*(5), 779–791. http://doi.org/10.3758/BF03194972

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Tsang, T., Atagi, N., & Johnson, S. P. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *Journal of Experimental Child Psychology*, *169*, 93–109. http://doi.org/10.1016/j.jecp.2018.01.002

van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *Journal of Speech, Language, and Hearing Research*, *55*(2), 554. http://doi.org/10.1044/1092-4388(2011/10-0347)

van Wassenhove, V., Grant, K. W., Poeppel, D., & Halle, M. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. http://doi.org/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. *Journal of Neuroscience*, *39*(15), 2938–2950. http://doi.org/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. *Perception & Psychophysics*, *60*(6), 926–940. http://doi.org/10.3758/bf03211929

Von Holzen, K., Nishibayashi, L., & Nazzi, T. (2018). Consonant and vowel processing in word form segmentation: An infant ERP study. *Brain Sciences*, *8*(2), 24. http://doi.org/10.3390/brainsci8020024

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344–356. http://doi.org/10.1016/j.wocn.2009.04.002

Weatherhead, D., & White, K. S. (2017). Read my lips: Visual speech influences word processing in infants. *Cognition*, *160*, 103–109. http://doi.org/10.1016/j.cognition.2017.01.002

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*, 173–196. http://doi.org/10.1146/annurev-psych-010814-015104

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*(5), 672–683. https://doi.org/10.1037/0012-1649.24.5.672

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. http://doi.org/10.1016/s0163-6383(84)80022-3

Werker, J. F., & Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology*, *46*(3), 233–251. http://doi.org/10.1002/dev.20060

Wickham, H., & Henry, L. (2020). tidyr: Tidy messy data [Computer software manual]. http://CRAN.R-project.org/package=tidyr (R package version 1.0.2)

Wickham, H., François, R., Henry, L. & Müller, K. (2020). dplyr: A grammar of data manipulation [Computer software manual]. https://CRAN.R-project.org/pacpkage=dplyr (R Package version 0.7.6.)

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *NeuroImage*, *172*, 766–774. http://doi.org/10.1016/j.neuroimage.2018.01.038

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

Yang, X., Wang, K., & Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, *38*(2), 824–839. http://doi.org/10.1109/18.119739

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171–211). Springer.

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

Yoon, J. M. D., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *105*(36), 13690–13695. http://doi.org/10.2307/25464105

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

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology*, *28*(3), 401–408. http://doi.org/10.1016/j.cub.2017.11.071

Appendix A: Information Sheets and Consent Forms

Seeds of Language Development

INFORMATION SHEET FOR RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants and Children with and without Hearing Impairment

Support for this study: HEARing Cooperative Research Centre
Western Sydney University

Invitation to Participate and Description of Project
You are invited to take part in a study to investigate infants' language development and hearing impairment. The study is funded by the HEARing Cooperative Research Centre and led by Professor Denis Burnham and Dr Marina Kalashnikova. The aim is to observe how young infants perceive speech as registered via the non-intrusive electroencephalogram (EEG) and eye-tracking procedures.

What does the study involve?
If you choose to take part in this study, you will first be asked to fill out a questionnaire about you and your child. In the study, your child will sit in a comfortable infant or child chair or on your lap in front of a video monitor. During the session, we will measure your child's brain waves and eye movements. Your child's brain waves will be recorded to a computer using an advanced EEG system and a sensor net. The smooth net of sensors will be placed on your child's head before testing begins. This only takes a few minutes, and most children quickly forget they are wearing it. Your child's eye movements will be recorded via a remote Tobii eye tracker system. Both EEG and Tobii eye tracker systems are certified for use with infants and children, and are used extensively in infant and child studies. These procedures are completely non-intrusive and painless and are regularly used with babies. The study will last around 40 minutes and will be concluded if your child becomes fussy, or if you wish to finish.

How much time will the study take?
The study is approximately 40 minutes in duration.

Will the study involve any discomfort for me or my child?
The study is not intended to involve any discomfort to you or your child. The only risk is that your child may experience some boredom or restlessness, in which case we would stop immediately.

Will the study benefit me?
There will not be any specific immediate benefits to any individual. But more general benefits of this research include: the possibility of identifying the aspects of language acquisition that predict later language delay associated with hearing loss; researchers developing new devices; clinicians developing new intervention and remediation strategies; and early intervention, education and clinical end-users providing (re)habilitation services.

Appendix A-1. *Information Sheet for infant and children participants in Study 1 (page 1 of 2)*

Seeds of Language Development

How is this study being paid for?
This study is supported by the HEARing Cooperative Research Centre and internal funding from Western Sydney University. You will be compensated $30 for your travel expenses at each visit.

Will anyone else know the results? How will the results be disseminated?
The research team will report the results at conferences and in academic journals. All information about the participants will remain confidential. If you would like to know about the results, we can send them to you once the study is completed.
Excerpts from the recordings may also be used for illustrative purposes in teaching, in conference presentations or in various relevant electronic media but neither you nor your child will be identified if the recordings are used for any of these purposes.

Confidentiality
Personal information gathered in the course of the study is confidential and will be securely stored. No personal information will be given to any persons other than the researchers unless it is made anonymous.

Can I withdraw from the study?
Participation is entirely voluntary: you are not obliged to be involved and, if you do participate, you can withdraw at any time without giving any reason and without any consequences.

Can I tell other people about the study?
Yes please, you can tell other people about the study and they can contact the BabyLab Coordinator (Jen Armstrong on 9772 6696) to register their interest.

What if I require further information?
When you have read this information, we will discuss it with you further and answer any questions you may have. Please feel free to ask about anything you don't understand and to consider this consent form for as long as you feel is necessary before you decide whether to participate.

If you would like to know more at any stage of the study, please contact us. We will be happy to discuss it with you:

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Dr Marina Kalashnikova, (02) 9772 6264; m.kalashnikova@westernsydney.edu.au
Dr Benjawan Kasisopa, (02) 9772 6269; b.kasisopa@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

What if I have a complaint?
This study has been approved by the University of Western Sydney Human Research Ethics Committee. The Approval number is H11517.

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0883 Fax 02 4736 0013 or email humanethics@westernsydeny.edu.au.

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix A-1. *Information Sheet for infant and children participants in Study 1 (page 2 of 2)*

Seeds of Language Development

the MARCS institute for brain, behaviour and development

THE HEARING CRC®

WESTERN SYDNEY UNIVERSITY

CONSENT FORM for RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants and Children with and without Hearing Impairment

1. I agree to take part in the Auditory-Visual Speech Perception study as described in the Information Sheet.

2. I have read the Information Sheet for Research Participants. This Information Sheet tells me what the research is for, and what participants do in the study.

3. I have had a chance to ask questions. I was given complete answers to my questions.

4. I understand that the researcher may contact the clinic for the hearing impairment record of my child.

5. I agree to release my child's hearing impairment clinical record to be kept in this project.

6. I can withdraw from the study at any time. If I withdraw from the study, I understand that Western Sydney University will not discriminate against me in the future.

7. The results can be published and the results can be presented. I will be identified only by an identification code. My name will not be used in research presentations or publications.

8. I agree to be contacted for participation in future research project. YES / NO

9. If I have questions, I can contact

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Dr Marina Kalashnikova, (02) 9772 6264; m.kalashnikova@westernsydney.edu.au
Dr Benjawan Kasisopa, (02) 9772 6269; b.kasisopa@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

10. I am keeping a copy of this information sheet and consent form.

Name (please PRINT) _____

Signature: _____

Date: _____

Appendix A-2. *Consent Form for infant and children participants in Study 1 (page 1 of 2)*

Seeds of Language Development

Signature of Person Obtaining Consent Phone

_____ _____

Note: This study has been approved by the Western Sydney University Human Research Ethics Committee. The Approval number is H11517. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0493 (Ext 2493) or email humanethics@westernsydeny.edu.au.
Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix A-2. *Consent Form for infant and children participants in Study 1 (page 2 of 2)*

Seeds of Language Development

INFORMATION SHEET FOR RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants, and Children and Adults with and without Hearing Impairment

Support for this study: HEARing Cooperative Research Centre
Western Sydney University

Invitation to Participate and Description of Project
You are invited to take part in a study to investigate language development and hearing impairment. The study is funded by the HEARing Cooperative Research Centre and led by Professor Denis Burnham. The aim is to observe how infants, children and adults perceive speech using the non-intrusive electroencephalogram (EEG) and eye-tracking procedures.

What does the study involve?
If you choose to take part in this study, you will first be asked to fill out a questionnaire about you. In the study, you will sit in a chair in front of a video monitor. During the session, we will measure your brain waves and eye movements. Your brain waves will be recorded using an advanced EEG system and a sensor net. The smooth net of sensors will be placed on your head before testing begins. Your eye movements will be recorded via a remote eye tracker system. Both the EEG and eye tracker systems are certified for use with infants, children and adults. These procedures are completely non-intrusive and painless and are regularly used.

How much time will the study take?
The study is approximately 40 minutes in duration.

Will the study involve any discomfort for me?
The study is not intended to involve any discomfort to you. The only risk is for you to experience some boredom or restlessness.

Will the study benefit me?
There will not be any specific immediate benefits to any individual. Benefits of this research include: the possibility of identifying the aspects of language acquisition that predict later language delay associated with hearing loss; researchers developing new devices; clinicians developing new intervention and remediation strategies; and early intervention, education and clinical end-users providing (re)habilitation services.

How is this study being paid for?
This study is supported by the HEARing Cooperative Research Centre and internal funding from Western Sydney University. You will be compensated 5 course credits for your participation.

Will anyone else know the results? How will the results be disseminated?
The research team will report the results at conferences and in academic journals. All information about the participants will remain confidential. If you would like to know about the results, we can

Appendix A-3. *Information Sheet for adult participants in Study 1 (page 1 of 2)*

send them to you once the study is completed.
Excerpts from the recordings may be also be used for illustrative purposes in teaching, in conference presentations or in various relevant electronic media but you will be identified if the recordings are used for any of these purposes.

Confidentiality
Personal information gathered in the course of the study is confidential and will be securely stored. No personal information will be given to any persons other than the researchers unless it is made anonymous.

Can I withdraw from the study?
Participation is entirely voluntary: you are not obliged to be involved and, if you do participate, you can withdraw at any time without giving any reason and without any consequences.

Can I tell other people about the study?
Yes please, you can tell other people about the study and they can contact the BabyLab Coordinator (Jen Armstrong on 9772 6696) to register their interest.

What if I require further information?
When you have read this information, we will discuss it with you further and answer any questions you may have. Please feel free to ask about anything you don't understand and to consider this consent form for as long as you feel is necessary before you decide whether to participate.

If you would like to know more at any stage of the study, please contact us. We will be happy to discuss it with you:

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

What if I have a complaint?
This study has been approved by the University of Western Sydney Human Research Ethics Committee. The Approval number is H11517.

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0883 Fax 02 4736 0013 or email humanethics@westernsydeny.edu.au.

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix A-3. *Information Sheet for adult participants in Study 1 (page 2 of 2)*

Seeds of Language Development

the MARCS institute for brain, behaviour and development

THE HEARING CRC®

WESTERN SYDNEY UNIVERSITY

CONSENT FORM for RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants, Children and Adults with and without Hearing Impairment

1. I agree to take part in the Auditory-Visual Speech Perception study as described in the Information Sheet.

2. I have read the Information Sheet for Research Participants. This Information Sheet tells me what the research is for, and what participants do in the study.

3. I have had a chance to ask questions. I was given complete answers to my questions.

6. I can withdraw from the study at any time. If I withdraw from the study, I understand that Western Sydney University will not discriminate against me in the future.

7. The results can be published and the results can be presented. I will be identified only by an identification code. My name will not be used in research presentations or publications.

8. I agree to be contacted for participation in future research projects. YES / NO

9. If I have questions, I can contact

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

10. I am keeping a copy of this information sheet and consent form.

Name (please PRINT) _____

Signature: _____

Date: _____

Appendix A-4. *Consent Form for adult participants in Study 1 (page 1 of 2)*

Seeds of Language Development

Signature of Person Obtaining Consent Phone

Note: This study has been approved by the Western Sydney University Human Research Ethics Committee. The Approval number is H11517. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0493 or email humanethics@westernsydeny.edu.au.
Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix A-4. *Consent Form for adult participants in Study 1 (page 2 of 2)*

Seeds of Language Development

the MARCS institute for brain, behaviour and development

THE HEARING CRC®

WESTERN SYDNEY UNIVERSITY

INFORMATION SHEET FOR RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants and Children with and without Hearing Impairment

Support for this study: HEARing Cooperative Research Centre
Western Sydney University

Invitation to Participate and Description of Project
You are invited to take part in a study to investigate infants' language development and hearing loss. The study is funded by the Hearing Cooperative Research Centre and led by Professor Denis Burnham and Dr Marina Kalashnikova. The aim of this study is to observe how young infants perceive different speech registered via an eye tracking procedure. This study will take place at the Shepherd Centre where your child has regular appointment.

What does the study involve?
If you choose to take part in this study, you will be asked to fill out a questionnaire about you and your child. Your child will sit in a comfortable chair or on your lap in front of a monitor. During the session, we will measure your child's eye movements and record them to a computer using a Tobii eye tracker system. This system is certified for use with infants and children, and it is used extensively in infant and child studies. These procedures are completely painless and are regularly used with babies. The study will last around 30 minutes and will be concluded if your child becomes fussy, or if you wish to finish.

How much time will the study take?
The study is approximately 30 minutes long.

Will the study involve any discomfort for me or my child?
The study is not intended to involve any discomfort to you or your child. The only risk is for your child to experience some boredom or restlessness, in which case we would stop immediately.

Will the study benefit me?
There will not be any specific immediate benefits to any individual. But more general benefits of this research include: the possibility of identifying the aspects of language acquisition that predict later language delay associated with hearing loss; researchers developing new devices; clinicians developing new intervention and remediation strategies; and early intervention, education and clinical end-users providing (re)habilitation services.

How is this study being paid for?
This study is supported by the HEARing Cooperative Research Centre (HEARing CRC) and internal funding from Western Sydney University. Your child will receive a small gift and a BabyLab scientific degree at the end of the study.

Appendix A-5. *Information Sheet for Study 2 (page 1 of 2)*

Seeds of Language Development

Will anyone else know the results? How will the results be disseminated?

The research team will report the results at conferences and in academic journals. All information about the participants will remain confidential. If you would like to know about the results, we can send them to you once the study is completed.

Excerpts from the recordings may also be used for illustrative purposes in teaching, in conference presentations or in various relevant electronic media but neither you nor your child will be identified if the recordings are used for any of these purposes.

Confidentiality

Personal information gathered in the course of the study is confidential and will be securely stored. No personal information will be given to any persons other than the researchers unless it is made anonymous.

Can I withdraw from the study?

Participation is entirely voluntary: you are not obliged to be involved and, if you do participate, you can withdraw at any time without giving any reason and without any consequences.

Can I tell other people about the study?

Yes please, you can tell other people about the study and they can contact the MARCS BabyLab (Jen Armstrong on 9772 6696) to register their interest.

What if I require further information?

When you have read this information, we will discuss it with you further and answer any questions you may have. Please feel free to ask about anything you don't understand and to consider this consent form for as long as you feel is necessary before you decide whether to participate.

If you would like to know more at any stage of the study, please contact us. We will be happy to discuss it with you:

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Dr Marina Kalashnikova, (02) 9772 6264; m.kalashnikova@westernsydney.edu.au
Dr Benjawan Kasisopa, (02) 9772 6269; b.kasisopa@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

What if I have a complaint?

This study has been approved by the Western Sydney University Human Research Ethics Committee. The Approval number is H11517.

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0883 Fax 02 4736 0013 or email HumanEthics@westernsydney.edu.au.

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix A-5. *Information Sheet for Study 2 (page 2 of 2)*

Seeds of Language Development

the MARCS institute for brain, behaviour and development

THE HEARING CRC®

WESTERN SYDNEY UNIVERSITY

CONSENT FORM for RESEARCH PARTICIPATION

Project Title: The Auditory-Visual Speech Perception of Infants and Children with and without Hearing Impairment

1. I agree to take part in the Auditory-Visual Speech Perception study as described in the Information Sheet.

2. I have read the Information Sheet for Research Participants. This Information Sheet tells me what the research is for, and what participants do in the study.

3. I have had a chance to ask questions. I was given complete answers to my questions.

4. I understand that the researcher may contact the clinic for the hearing loss record of my child.

5. I agree to release my child's hearing loss clinical record to be kept in this project.

6. I can withdraw from the study at any time. If I withdraw from the study, I understand that Western Sydney University will not discriminate against me in the future.

7. The results can be published and the results can be presented. I will be identified only by an identification code. My name will not be used in research presentations or publications.

8. I agree to be contacted for participation in future research project. YES / NO

9. If I have questions, I can contact

Professor Denis Burnham, (02) 9772 6677; d.burnham@westernsydney.edu.au
Dr Marina Kalashnikova, (02) 9772 6264; m.kalashnikova@westernsydney.edu.au
Dr Benjawan Kasisopa, (02) 9772 6269; b.kasisopa@westernsydney.edu.au
Jessica Tan, (02) 9772 6535; j.tan@westernsydney.edu.au

10. I am keeping a copy of this information sheet and consent form.

Name (please PRINT) _____

Signature: _____

Date: _____

Appendix A-6. *Consent Form for Study 2 (page 1 of 2)*

Seeds of Language Development

Signature of Person Obtaining Consent Phone

Note: This study has been approved by the Western Sydney University Human Research Ethics Committee. The Approval number is H11517. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel 02 4736 0883 Fax 02 4736 0013 or email HumanEthics@westernsydney.edu.au.
Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

4 of 3

Appendix A-6. *Consent Form for experiment presented in Chapter 4 (page 2 of 2)*

Appendix A-6. *Consent Form for Study 2 (page 2 of 2)*

Appendix B: Questionnaire

BABYLAB

Infant's Name:.....

MARCS BABY LAB ***Family Information Sheet***

(All information is strictly confidential. Questions marked with an asterisk* are optional)

Infant's date of birth: _____ Mother's Age (today):_____ Father/ Partner's Age (today): _____

* Mother's Name: _____* Father/ Partner's Name: _____

* Mother's Occupation: _____ * Father/ Partner's Occupation: _____

* Mother's Education: (Please circle both secondary and tertiary level completed)

Secondary Education; Year 10 Year 11 Year 12

Tertiary Education; TAFE University Masters Ph.D. Other_____

*Father/ Partner's Education: (Please circle both secondary and tertiary level completed)

Secondary Education; Year 10 Year 11 Year 12

Tertiary Education; TAFE University Masters Ph.D. Other_____

1. Were there any complications of Pregnancy _____
   and/or Labour/Delivery _____
2. Have you ever been diagnosed with PND? Yes ☐ No ☐
3. Was your infant: Fullterm 38-42 weeks ☐ Premature ≤ 37 weeks ☐.....weeks Post-mature >42 weeks ☐
4. What was your infant's (a) Birthweight? _____ kg (b) Apgar score? _____ (0-10)
5. Did your infant complete the newborn hearing screen? Passed ☐ Concerns ☐ _____
6. Do you have any concerns about your infant's hearing? Yes ☐ No ☐ Please describe: _____
   _____
7. Has your infant had any medical/other problems? Yes ☐ No ☐ Please describe:_____
   _____
8. In your child's family is there a history of:
   (a) Hearing impairment/deafness Yes ☐ No ☐
   If yes, relation to child? _____Type/Degree?_____
   (b) Reading, speech, and/or language problems in the family (i.e. Dyslexia) Yes ☐ No ☐
   If yes, relation to child? _____Type/Degree?_____
9. What is the primary language spoken in your home? _____

Mother's first language _____ Partner's first language _____

(a) Please list any other languages/accents (including English accents) that are spoken in your home:

1. _____ Hours/week spoken?_____
2. _____ Hours/week spoken?_____
3. _____ Hours/week spoken?_____

Office Use Only: Date: ..... ID Number: .....

Appendix B. *Participant Information Questionnaire (page 1 of 2)*

BABYLAB

Infant's Name:.....

10. How is your baby's health today?

1st Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

2nd Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

3rd Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

4th Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

5th Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

6th Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

7th Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

8th Visit:_____[Date: __/__/__] Recent Ear Infection? Y ☐ How long ago?.....

11. Has your infant ever had an ear infection? Yes ☐ No ☐ If yes, how many? _____

You can help Baby Science! Even your $30 can make a huge difference.

If you would like to help Baby Science further and donate your $30 travel money or a portion of your travel money, please indicate below:

1st Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
2nd Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
3rd Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
4th Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
5th Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
6th Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
7th Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐
8th Visit:	$30 ☐	$20 ☐	$10 ☐	Not this time ☐

THANK YOU FOR YOUR SUPPORT!

Office Use Only: Date: ..... ID Number: .....

Appendix B. *Participant Information Questionnaire (page 2 of 2)*

Appendix C: Stimuli used in Study 1 (Chapter 3)

1.	Hi baby! How are you today? It's a wonderful day today! You look happy! Are you ready for some fun?
2.	We're going to read a story now! Do you want to read a story? This story is about a sheep, shoe and shark. Here we go!
3.	Here comes the sheep! It's a lovely sheep! It's nice and fluffy! The sheep has white wool! Can you feel the soft wool?
4.	Does the sheep have a tail? Does the sheep have ears? Can you see the two ears? The nice soft sheep has two ears!
5.	This sheep has a black nose! What noise does the sheep make? Ba-a ba-a. You like the sheep, don't you?
6.	Let's have a look at some pictures! What have we got here? This is a funny picture! Should we look at another picture? Look at this one!
7.	What is this? This is a very nice shoe! What colour is the shoe? It's a red shoe. I like red shoes!
8.	Shoes go on feet! Can we put the shoes on your feet? Do you like the shoes? Do you want to hold the shoes?
9.	These are some beautiful flowers! That's a daisy! There's a butterfly on the daisy! Can you see the butterfly?
10.	Can you see the sky? Look at the blue sky! Isn't it beautiful? The clouds look like fairy floss!
11.	Uh oh! Someone threw the shoe up into the tree! Isn't that silly? What should we do now?
12.	Can you look here? Here is the shark! Can you see the shark? The shark is having a good time!
13.	Would you like to clap for me? Clap clap clap! Clap your hands! You love that, don't you?
14.	It's a beautiful day today! I see the sun shining brightly! It's really nice that you came to see me today!

15.	You came from far away to see me today! It's really nice to meet you! I hope it's also the case for you! You are a beautiful baby!
16.	We're going to the park later! The weather is great for some outdoor time! I bet we're going to see many dogs! What do you think?
17.	My! Look at those shoes! There are pink flowers on them! Aren't they pretty? I think I can wear them all day! Do you like them too?
18.	Look! I have four soft toys. How many have you got? My favourite is this brown teddy bear! I take it everywhere with me!
19.	What would you like to do today? It's a very sunny day! Shall we go to the beach? Let's head to the beach! It's time to find our cossies!
20.	There's a little bird over there! It's green and red and yellow! Oh, look! It's eating a worm! The little bird must be hungry!
21.	A rainbow has seven colours! Red, orange, yellow, green, blue, violet, and indigo. My favourite colour is blue! What's your favourite colour?
22.	Can you hear that sound? I wonder where it's coming from! Oh, it's coming from under that box! Will we take a look?
23.	Whee! This is fun! I love playgrounds! My favourite part is the slide! The higher the slide, the better it is! Do you like slides too?
24.	Look! A kitten's coming my way! It has grey patches all over it! It's so cute! I want to play with it! Let's go!
25.	It's me again! Look at what I've got! What do we have here? What's in here? Oh, it's an apple! I love eating apples!
26.	Wow! Look at that tree! It has lots of pretty pink flowers on it! Let's go closer so we can have a better look!
27.	I like going to parks! There's such a huge space to run about! What about you? Do you like parks too?
28.	When the weather gets too hot, I like to wear my cap! The cap keeps the sun off my face. Do you also wear a cap?
29.	When the weather gets too cold, I always drink hot chocolate! It keeps me warm! What's your favourite drink?
30.	You've been so attentive today! Thank you for listening to me! I enjoyed talking to you! I hope you enjoyed listening to me too!

Appendix D: Stimuli used in Study 2 (Chapter 4)

Taken from Jusczyk and Aslin (1995).

Target word	Six-sentence passages
Cup	The cup was bright and shiny.
	A clown drank from the red cup.
	The other one picked up the big cup.
	His cup was filled with milk.
	Meg put her cup back on the table.
	Some milk from your cup spilled on the rug.
Dog	The dog ran around the yard.
	The mailman called to the big dog.
	He patted his dog on the head.
	The happy red dog was very friendly.
	The dog barked only at squirrels.
	The neighborhood kids played with your dog.
Feet	The feet were all different sizes.
	This girl has very big feet.
	Even the toes on her feet are large.
	The shoes gave the man red feet.
	His feet get sore from standing all day.
	The doctor wants your feet to be clean.

Bike	His bike had big black wheels.
	The girl rode her big bike.
	Her bike could go very fast.
	The bell on the bike was really loud.
	The boy had a new red bike.
	Your bike always stays in the garage.

Appendix D. *Stimuli used in Study 2 (page 2 of 2)*