

**Radboud University**



Donders Graduate School

Master of Science Programme Cognitive Neuroscience (Research)

MSc Thesis

**Audio-visual cues in language development: Do visual  
speech cues enhance infants' cortical speech tracking?**

Antonia Jordan Monteiro de Barros

Supervisors: Melis Çetinçelik and Tineke Snijders

Second reader: Sabine Hunnius

Date of final oral examination: 17<sup>th</sup> March 2022

Radboud University Nijmegen



MAX PLANCK INSTITUTE  
FOR PSYCHOLINGUISTICS

## Abstract

Infants are sensitive to audio-visual speech cues from a very young age. In this study, we investigated the role of visual speech cues, such as a speaker's rhythmic movements of the mouth, lips and jaw, on infants' cortical speech tracking. In adults, research has shown that seeing congruent audio-visual speech enhances neuronal tracking of the speech envelope. This is found specifically between 2-7 Hz which corresponds to the rate in which syllables appear in adult-directed speech and thus the frequencies in which mouth movements are tightly locked to the speech envelope. Using electroencephalography (EEG), we investigated whether 10-month-old Dutch infants (N=7) also show enhanced cortical speech tracking when a speaker's visual speech cues are visible compared to a condition in which the speaker's mouth and jaw movements are occluded with a block. We predicted that visual speech cues would enhance infants' cortical speech tracking, specifically at the syllable rate (2.5–3.5 Hz for our stimuli) and stressed syllable rate (1–1.75 Hz). First, our results show that infants looked significantly longer towards audio-visual stimuli compared to stimuli in which the speaker's visual speech cues were occluded. Furthermore, our findings suggest that visual speech cues indeed enhanced infants' cortical speech tracking. However, this was not found at the syllable or the stressed syllable rate, but instead at 3.75–4 Hz, a frequency range that corresponds to the theta band in infants. Spectral power of the EEG signal was also found to be enhanced by the presence of visual speech cues at the theta range (3.75–4 Hz). Our results suggest that theta-band oscillations may play a role in merging multi-modal information, such as the visual speech signal and the speech envelope. Furthermore, our findings provide further evidence for infants' sensitivity to audio-visual cues and highlight their influence on infants' language processing.

## **1. Introduction**

Speech perception is commonly regarded and studied as a strictly auditory event. However, speech is most often produced in social-interactive settings and is thus accompanied by many visual cues, making it inherently multi-modal. This is also the case for language acquisition. Infants are typically not exposed to speech in isolation, but in combination with facial cues of their caregivers during face-to-face interactions. These facial cues include the speaker's rhythmic movements of the mouth, lips and jaw during speech production. These movements are also referred to as visual speech cues or articulatory cues and they often occur in synchrony with the sounds that are produced. In adults, seeing the correspondence between the articulators and the incoming speech sounds has long been known to facilitate speech perception (Erber, 1975). This is especially the case when speech is degraded, such as in a noisy environment or when listening to a speaker with a strong foreign accent (Sumbly & Pollack, 1954; Zheng & Samuel, 2019). For infants, who are just starting to tune into the sound system of their native language, speech perception poses a challenge even in optimal listening conditions. Therefore, audio-visual speech cues may be even more beneficial for infants than adults.

### **1.1. The role of audio-visual cues in early language development**

Research suggests that infants are sensitive to visual cues and integrate visual speech information with the auditory stream from a very young age. For instance, studies have shown that two-month-old infants already look longer towards a face articulating the sound they hear as opposed to a face articulating a mismatched sound (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003). Moreover, the McGurk effect (McGurk & MacDonald, 1976), an illusory effect of integrating incongruent auditory and visual information during speech perception, has been found in infants as young as four months of age (Burnham & Dodd, 2004). Neuroimaging studies further illustrate infants' audio-visual speech integration. For instance, Kushnerenko et al. (2008) presented five-month-old infants with incongruent audio-visual information that could either be fused into a new phoneme (visual /ga/ matched with auditory /ba/ is perceived as /da/) or incongruent information that is non-fusable for adults (visual /ba/ with auditory /ga/ is perceived as

/bga/). Measuring infants' event-related brain potentials (ERPs), they found a mismatching brain response only when the conflicting information led to an illegal phoneme (/bga/), but not when integration of the two modalities led to a fusible phoneme (/da/). This demonstrates that, at five months, infants already show an adult-like ability to perceive and integrate visual speech cues with auditory information.

Furthermore, Teinonen et al. (2008) demonstrated that six-month-old infants learn phonemic boundaries depending on the visual articulation they were exposed to. In their study, infants were familiarised to audio-visual stimuli ranging on an acoustic continuum between /ba/ and /da/. One group of infants (the one-category condition) was exposed to all tokens in combination with the visual articulation of either /ba/ or /da/. Conversely, the group in the two-category condition was exposed to the tokens that are typically perceived by adult listeners as /ba/ onto a visual articulation of /ba/ and all tokens that are perceived as /da/ onto a visual articulation of /da/. Thus, the two groups were exposed to different distributions of visual articulation, but the acoustic stimuli they were exposed to was identical. Crucially, the results show that only the infants in the two-category condition were able to distinguish the /ba-/da/ contrast while infants in the one-category condition did not discriminate between the two phonemes. This provides strong evidence that visual information can actively influence infants' acquisition of phonemic boundaries. Overall, these studies suggest that infants are able to match visual information from the articulators with the acoustic stimuli they encounter and may use this cross-modal relationship to acquire language.

More specifically, visual cues provided by the mouth may be especially informative for infants. Attention to the mouth at six and at twelve months has been found to significantly predict expressive language abilities at 18 months (Young et al., 2009; Tenenbaum et al., 2015). Interestingly, there seems to be a developmental stage towards the end of the first year during which infants shift their attention from the speaker's eyes to the speaker's mouth. Tenenbaum et al. (2013) found that this shift was strongest between nine and twelve months, but other studies indicate that attention to the mouth already gradually increases between four and nine months of age for both mono- and bilingual infants (Mercure et al., 2019). Similarly, Lewkowicz and Hansen-Tift (2012) found that infants look longer at the speaker's mouth compared to the eyes at eight and ten

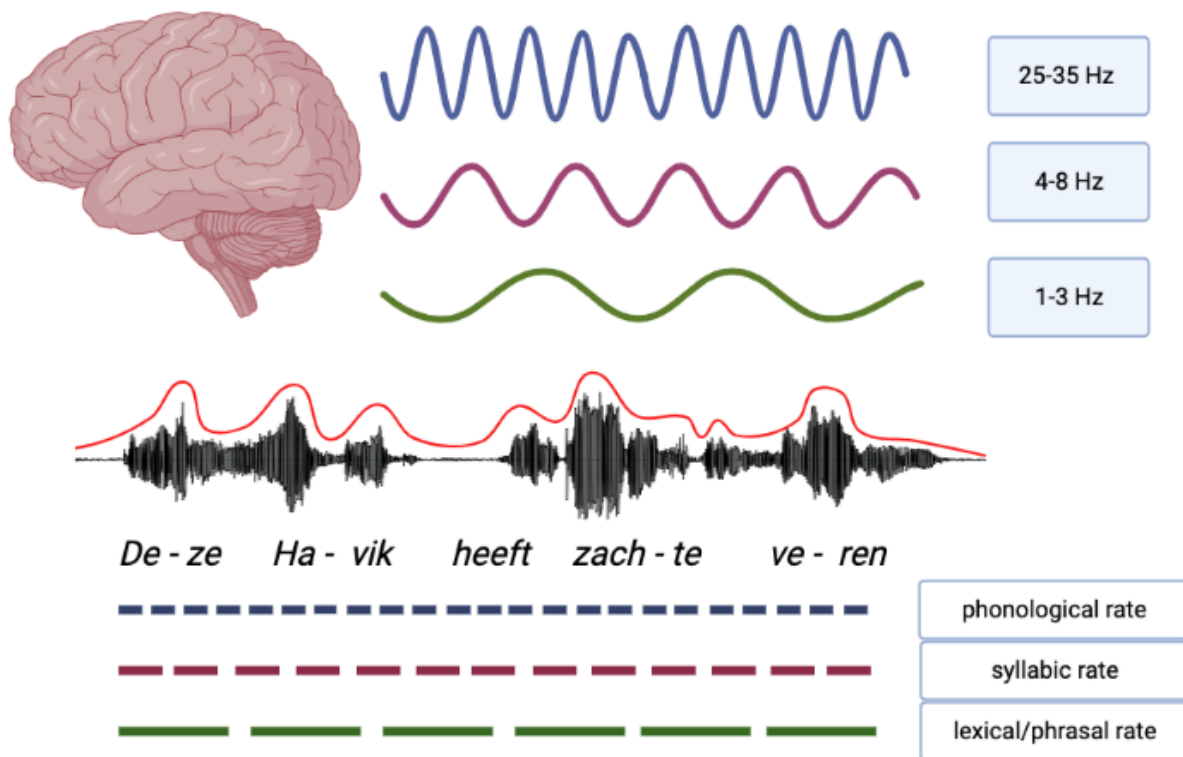
months, and that they began shifting their attention back towards the eyes at twelve months for their native language, but kept their attention towards the mouth for non-native languages. Overall, these studies demonstrate that infants start shifting their gaze away from the eyes at four months of age, but only look more at the mouth than the eyes between eight and twelve months. These findings are especially intriguing considering that this developmental stage reflects a sensitive period in infants' language acquisition, marking the onset of canonical babbling and infants' phonetic attunement into the specific sound system of their native language (Kuhl et al., 2006). These studies propose that when infants start acquiring the specific properties of their respective languages, visual cues provided by the mouth may become as important as the social cues provided by the eyes. In summary, infants appear to be sensitive to audio-visual cues from an early stage and this sensitivity may be beneficial to their future language development.

## **1.2. The role of audio-visual cues on cortical speech tracking**

Although research suggests that infants pay attention towards the speaker's articulators, it remains unclear whether visual speech cues can facilitate infants' on-line speech processing, and if yes, through which mechanisms this facilitation occurs. One proposed mechanism for successful speech processing that may be modulated by the presence of audio-visual cues is cortical speech tracking (Luo et al., 2010; Poeppel, 2014).

Cortical speech tracking refers to the process by which neural oscillations synchronise with the incoming speech signal. This synchronisation usually occurs through aligning the phase or amplitude of the ongoing oscillatory activity to the frequencies present in the speech envelope (Giraud & Poeppel, 2012; Peelle & Davis, 2012; Di Liberto et al., 2015). Cortical speech tracking, sometimes referred to as neural entrainment (although the terminology is debated - see Haegens (2020)), is proposed to be an important mechanism for efficient speech processing (Peelle & Davis, 2012). By resetting the phase of oscillatory activity to salient features in the speech envelope, important information in the acoustic signal is able to arrive at a point of high neuronal excitability. The idea is that this phase-locking facilitates parsing of continuous speech into meaningful linguistic units (Giraud & Poeppel, 2012; Ding et al., 2016). These linguistic units (e.g., syllables, words, phrases) appear at different rates in the speech stimuli that correspond

to different frequency bands in the neural signal. In particular, in typical adult-directed speech (ADS), the speech waveform carries linguistic information at the following timescales (Giraud & Poeppel, 2012): the phonological rate (25-35 Hz) which corresponds to the low gamma range; the syllabic rate (4-8 Hz) corresponding to the theta range; and the lexical/phrasal rate (1-3 Hz) which denotes the delta range (see Figure 1). However, for infant-directed speech (IDS), which is typically characterised by a slower speech rate, these frequencies are often lower.



**Figure 1.** Cortical speech tracking occurs when ongoing neural oscillations entrain to external stimuli, such as the speech signal. This entrainment occurs at different neural frequencies (top) which correspond to the frequencies of linguistic units present in the speech envelope (bottom). Frequencies given here are for ADS. Figure created with [BioRender.com](https://www.biorender.com).

The vast majority of studies on cortical speech tracking have been conducted on adults. Thus, the influence of neuronal speech entrainment on infants' language acquisition is not yet clear. If cortical speech tracking is indeed linked to successful speech encoding, then understanding its role on language acquisition is of key interest to the field

of child language development as it may be a mechanism that helps children extract information from the rapid and complex speech signal. To date, there are only few studies that investigated infants' cortical speech tracking. For instance, Kalashnikova et al. (2018) found that IDS enhances cortical tracking of speech compared to ADS in pre-verbal infants. The authors suggest that this may be due to the high salience of IDS, as its rhythmic features and exaggerated phonetic patterns could facilitate the synchronisation between the speech input and neural activity.

Furthermore, some studies have provided evidence that cortical speech tracking abilities may be related to other language outcomes. For instance, Snijders (2020) found that infants at 7.5 months show entrainment to both linguistic (infant-directed nursery rhymes) and non-linguistic (auditory beeps) stimuli. However, only infants' entrainment to linguistic stimuli was correlated with their word segmentation abilities at nine months. In particular, this correlation was found at the stressed syllable rate. These findings support the idea that cortical speech tracking may have functional relevance, such that it may help the listener parse and segment the incoming signal. Lastly, a recent study has shown that new-borns already show amplitude and phase tracking for both their native and non-native languages (Barajas et al., 2021). In summary, these first studies suggest that cortical speech tracking may be a mechanism that is present from birth and not necessarily dependent on speech comprehension, but that it may predict future language abilities.

In adults, one of the mechanisms argued to facilitate cortical speech tracking is access to visual speech cues (Luo et al., 2010; Crosse et al., 2015; Zoefel, 2021). These rhythmic cues of the mouth and lips provide the listener with additional information about the acoustic features of the incoming speech, such as the amplitude envelope or place and manner of articulation. In natural speech, articulatory movements can also provide temporal cues for upcoming syllables. Thus, the articulatory movements of visual speech may facilitate parsing of syllabic boundaries (Myers et al., 2019).

In particular, it is known that mouth movements are tightly synchronised to the speech envelope, meaning that there is a temporal correspondence between the opening and closing of the lips and the peaks and valleys of the speech envelope (Chandrasekaran et al., 2009). This was found particularly in the 2–7 Hz range (i.e., the syllable rate in ADS), suggesting that articulatory movements roughly occur in the same frequency as syllables

appear in speech. Therefore, having the rhythmic articulatory movements available during on-line speech processing may enhance speech perception and help the listener track the speech envelope, particularly at the syllable rate (Peelle & Sommers, 2015).

Indeed, research suggests that cortical speech tracking is enhanced when audio-visual cues are presented in both adverse and noise-free conditions (Crosse et al., 2015; 2016). In particular, Crosse et al. (2015) found that congruent visual information (i.e., visual speech that shares the timing of the acoustic signal) facilitates neural tracking of the speech envelope in adults. They investigated this by using natural continuous audio-visual speech and found that cortical speech tracking was higher for audio-visual than auditory-only speech, as well as higher than what an additive model of entrainment to the visual and auditory input separately would predict. This was found especially for the 2-6 Hz range, meaning that congruent audio-visual speech enhances cortical speech tracking especially at the syllable and stressed syllable rate. Furthermore, articulatory movements precede auditory information by about 100-300 ms during speech production (Chandrasekaran et al., 2009). It has thus been speculated that seeing the onset of articulation (e.g., the mouth beginning to open) may elicit a phase reset in the ongoing oscillatory activity, allowing the oscillations to phase-lock to the incoming speech stimuli through a predictive mechanism (Peelle & Sommers, 2015; Crosse et al., 2015, Zoefel, 2021). Evidence for this cross-modal phase modulation during audio-visual speech perception has been found in non-human primates (Lakatos et al., 2007; Kayser et al., 2008) as well as human adults (Micheli et al., 2020). For instance, using magnetoencephalography (MEG), Luo et al. (2010) found that the auditory cortex tracks both auditory and visual stimuli using low-frequency oscillations. Critically, they found that visual speech information present in congruent audio-visual speech reliably modulated the phase of the cortical oscillations in the auditory cortex (Luo et al., 2010), leading to more optimal cortical speech tracking. This was found once again at frequencies between 2–7 Hz. Overall, it appears that during congruent audio-visual speech perception, the rhythmic movements of articulators can be used to phase-lock the cortical activity in the auditory cortex to the speech envelope, enhancing cortical speech tracking.

So far, only few studies have investigated the influence of audio-visual cues on cortical speech tracking in children. Using auditory, visual, and audio-visual recordings of

a speaker producing the syllable /ba/, Power et al. (2012) investigated the role of visual cues on neural entrainment in 13-year-old children. While they found that entrainment was equal across all conditions in the delta range (2 Hz), they found a difference in entrainment for the audio-visual speech condition in the theta band (4 Hz). More specifically, they found that congruent visual speech cues modulate the phase of auditory entrainment in the theta band, converging with the findings reported in the adult studies above (Luo et al., 2010; Crosse et al., 2015). On the other hand, Tan et al. (2019) presented four-year-old children with naturalistic auditory, visual, and audio-visual materials while measuring their cortical tracking of speech. Here, they found no difference between the audio-visual and audio-only conditions, meaning that audio-visual speech did not enhance children's entrainment to the stimuli. It is unclear whether the difference in findings between the studies is due to developmental changes (i.e., 13-year-old children may show more adult-like processing of audio-visual speech compared to four-year-old children), or methodological differences such as the use of naturalistic vs. repetitive materials or the measure of analysis for neural entrainment. Nonetheless, despite the vast literature on the cross-modal relationship between visual and auditory stimuli in adult cortical speech tracking, no studies have tested this in infants yet.

### **1.3. The role of audio-visual cues on word segmentation**

Besides facilitating neural tracking of the speech envelope, some studies have shown that audio-visual cues may also facilitate word segmentation. Word segmentation is a crucial challenge that infants face when they begin to acquire language. In the vast majority of utterances, single words are not produced with periodic stops between them, but are most often embedded in a continuous stream of sounds. It has been shown that infants use statistical regularities and phonotactic cues (e.g., metrical stress) to segment speech, allowing them to recognize and learn their first words (Brooks & Kempe, 2012, Chapter 2).

In adults, studies have demonstrated that word segmentation is facilitated when synchronised visual speech cues are presented with the auditory stimuli. For instance, Mitchel & Weiss (2014) presented adults with an artificial language stream that contained almost no auditory cues (such as stress) about word boundaries. Participants were

exposed to audio-only and audio-visual materials where the speaker's facial movements were either informative (i.e., where their articulatory movements matched the word boundaries in the speech stream) or non-informative (i.e., where their facial movements did not match the word boundaries). Results showed that participants were better at segmenting the speech stream when visual information was available compared to the audio-only condition. However, this benefit was only found when the facial cues of the speaker were informative to the word boundaries. The study demonstrates that visual cues to word boundaries can influence auditory speech segmentation. Nonetheless, it is unclear which audio-visual cues specifically were employed for successful word segmentation (e.g., the rhythmic movements of the head, eye brows, lips or cheeks). Therefore, in a follow-up study, Lusk & Mitchel (2016) used eye-tracking to follow adults' gaze patterns when presented with a novel artificial language stream. Their results indicated that attention to the mouth in particular was related to speech segmentation, as adults showed longer looking times to the mouth compared to the eyes and nose during exposure to the artificial language stream. The authors suggest that the rhythmic opening and closing of the mouth provides the listener with reliable cues that mark syllabic boundaries. This can then enhance the segmentation of words and syllables in the continuous speech stream.

If this is indeed the case, then audio-visual cues provided by the articulators may be even more beneficial to infants as they are just tuning into the speech stream. Based on the adult studies mentioned above, it is possible that infants also use visual cues of articulatory movements to segment the speech stream and recognize words. As studies have shown that infants pay more attention to the mouth from around eight months of age, it is plausible that infants may use the rhythmic movements of the mouth and jaw as a cue for syllabic boundaries at that age. One study found that 7.5-month-old infants show better speech segmentation when they are able to see the synchronised face of the speaker compared to an unsynchronized or static face (Hollich et al., 2005). The authors also found that infants show better speech segmentation when presented with an oscilloscope pattern that moved in synchrony to the auditory stimuli, suggesting that it is not a moving face per se that aids segmentation, but any visual stimuli that is rhythmically congruent with the auditory input. However, the segmentation task in this study was performed in a

noise condition with a distractor voice being presented throughout the entire experiment. Although the adult studies provide evidence that visual cues can facilitate speech segmentation even in the absence of adverse listening conditions, infants' use of visual cues during speech segmentation in optimal (i.e., noise-free) conditions has, to our knowledge, not been tested so far.

#### **1.4. The current study**

Building on the previous literature, the aim of this study was to gain further insight into the multi-modal nature of language acquisition by investigating the role of visual speech cues on infants' language processing. While the relationship between infants' sensitivity to articulatory cues and their early phonetic learning is relatively well-established, no studies so far have investigated the role of audio-visual cues on infants' speech processing beyond that. Therefore, this study is part of a larger on-going project investigating the influence of visual speech cues on cortical speech tracking and word recognition in 10-month-old infants. Here, we present preliminary findings on the role of visual speech cues on infants' cortical speech tracking. With the current sample size of seven infants, however, we did not have enough statistical power to analyse the influence of visual speech cues on infants' word recognition. Thus, the results and discussion of this thesis pertain only to the role of audio-visual cues on infants' cortical speech tracking and not on their word recognition.

In this study, infants were presented with video materials in which the speaker's face was fully visible and video materials in which the speaker's visual speech cues (such as the movements of mouth, lips, cheeks and jaw) were occluded with a block. Electroencephalography (EEG) data was collected throughout the experiment. EEG data was used to answer our main research question, namely whether seeing the speaker's visual speech cues would enhance infants' cortical speech tracking (measured as speech-brain coherence). EEG data was also used to investigate whether visual speech cues would influence infants' spectral power at different frequency ranges. Furthermore, experimental sessions were video recorded for manual coding of the infants' looking times towards the screen. This was done in order to test whether looking times, and thus the infants' attention towards the stimuli, differed between the two experimental conditions.

We predicted that infants' speech-brain coherence would be enhanced when the speaker's articulatory cues were visible compared to the condition in which these cues were blocked. More specifically, we predicted that visual speech cues would enhance infants' speech-brain coherence at the syllable and the stressed syllable rate, as these bands were found to be significantly modulated by visual speech cues in the adult literature (Chandrasekaran et al., 2009; Luo et al., 2010; Crosse et al., 2015). Our research question was investigated by testing 10-month-old infants as this coincides with the developmental stage at which infants show the strongest sensitivity to audio-visual speech cues.

## **2. Methods**

### **2.1. Participants**

Seven Dutch-learning infants participated in the study (three male). Four other infants participated in the study, but were excluded due to technical error (N=2) or not contributing enough trials for the analysis (N=2). The mean age of participants was 312 days (range: 300-318 days), or 10 months and 7 days. Infants were eligible to participate if they were between 9.5-10.5 months of age and were raised in monolingual Dutch-speaking families. All participants were born full-term and had no developmental disability, no hearing or visual impairment, and no family history of language disorders.

Participants were recruited through the participant database of the Baby and Child Research Center, Radboud University. Caregivers had the choice between a book or 20 Euros as reimbursement for their participation. The experiment was approved by the Ethical Board of the Faculty of Social Sciences, Radboud University.

### **2.2. Materials**

Materials consisted of 240 sentences and 60 target words spoken by a female Dutch native speaker. The materials were created by selecting 30 target word pairs (i.e., 60 words in total) that were semantically related from the CELEX database (Baayen et al., 1995). As the target words were supposed to be unknown to the infants for the word segmentation task, only items with a low lexical frequency (< 3.89 per million in the CELEX database) were selected. Each target word was then embedded into four carrier sentences to create an experimental block. The semantically-related word pairs were embedded into the same carrier sentences to create A and B versions of the stimuli (see Table 1). The carrier sentences were 8-12 syllables long and the target words were bisyllabic trochaic Dutch words.

**Table 1.** Example of the A and B sentence versions presented during the familiarisation phase and their English translations.

Version A	Version B
<p><b>Ik doe altijd wat <b>dille</b> op mijn vis.</b>  <i>I always put some <b>dill</b> on my fish.</i></p>	<p><b>Ik doe altijd wat <b>venkel</b> op mijn vis.</b>  <i>I always put some <b>fennel</b> on my fish.</i></p>
<p><b>Duitse koks koken graag met <b>dille</b>.</b>  <i>German chefs like to cook with <b>dill</b>.</i></p>	<p><b>Duitse koks koken graag met <b>venkel</b>.</b>  <i>German chefs like to cook with <b>fennel</b>.</i></p>
<p><b>Wij hebben <b>dille</b> in de tuin.</b>  <i>We have <b>dill</b> in the garden.</i></p>	<p><b>Wij hebben <b>venkel</b> in de tuin.</b>  <i>We have <b>fennel</b> in the garden.</i></p>
<p><b>Zullen we in de salade wat <b>dille</b> doen?</b>  <i>Shall we put some <b>dill</b> in the salad?</i></p>	<p><b>Zullen we in de salade wat <b>venkel</b> doen?</b>  <i>Shall we put some <b>fennel</b> in the salad?</i></p>

One experimental block consisted of a familiarisation phase and test phase. The test phase was designed specifically to investigate infants' word recognition (data not reported here) whereas the familiarisation phase was used to analyse cortical speech tracking. In the familiarisation phase, infants saw a video of the female speaker producing the four sentences with the repeated target word. In the test phase, presented as audio-only, infants were exposed to two single words: the familiarised target and a novel control word. The novel word was always the corresponding word pair to the target word. Both the familiarisation and test phase materials were created using the same speaker and are spoken in IDS.

The video materials presented during the familiarisation phase showed the speaker's face, shoulders and chest with the speaker placed in front of a dark background. Video materials had a resolution of 680 x 680 and a frame rate of 50. Two versions of each video were created. In the audio-visual version, the speaker's face and accompanying visual speech cues were fully visible (AV condition). In the other version, a grey block was added in front of the speaker's lower face to occlude the speaker's articulatory movements (AV-Blocked condition). More specifically, the speaker's cheeks, mouth, jaw and larynx were blocked as they move in a rhythmic manner that is time-locked to the speech envelope (Chandrasekaran et al., 2009). The speaker's eyes remained

visible at all times and she kept direct eye gaze towards the camera. Figure 2 provides a visualization of both conditions.



**Figure 2.** Example of the two conditions: articulatory movements are visible on the left (AV condition) while they are hidden behind an occluder on the right (AV-Blocked condition).

The grey block was placed on top of the original videos using the DaVinci Resolve software (version 16; Black Magic Design, 2020). The block was static, meaning that it did not move during the video. However, as the speaker's head position differed between videos, the height of the block was adjusted for each individual file to occlude as many articulatory movements as possible without occluding her eyes. The width of the block remained the same across all videos. As the AV-Blocked videos were created by adding an occluder to the original AV videos, the auditory materials presented in both experimental conditions (AV vs. AV-Blocked) were completely identical.

Besides the videos presented in the familiarisation phase, nine attention-getters were used in this study to maintain infants' attention throughout the experiment. All of them were baby-friendly moving images placed over a black background with a repetitive sound effect.

Acoustic materials were normalized to 70 db using Praat (version 5.1.; Boersma & Weenink, 2014). Furthermore, Praat was used to determine the onset and duration of the familiarization and test phase. Praat was also used to identify the frequencies of the syllable and stressed syllable rate present in our stimuli. This was done by counting the number of syllables and stressed syllables present in each video and dividing those numbers by the duration of the overall speaking time in that video (excluding any pauses

at the start and end of the video as well as any pauses between sentences). This allowed us to measure the number of syllables and stressed syllables that appeared per second in our stimuli, providing us with the frequency bands of the syllable and stressed syllable rate. This revealed that, in our materials, syllables occurred at a rate of 2.6–3.6 Hz ( $M = 3.1$  Hz) and stressed syllables occurred at a rate of 0.9–1.7 Hz ( $M = 1.3$  Hz). Thus, the following bands were selected for subsequent analyses: 2.5–3.5 Hz for the syllable rate and 1–1.75 Hz for the stressed syllable rate. Lastly, acoustic properties of the stimuli, specifically duration, pitch and intensity, were compared between the A and B versions. None of the measures differed significantly between versions (see supplementary materials for mean and standard deviation values as well as statistical test results). This confirmed that materials in the A and B versions had comparable acoustic characteristics.

### **2.3. Design**

In total, eight different experimental lists were created. Half of the lists contained the target words from the A version, the other half contained the B versions. Furthermore, the order of appearance was counterbalanced for the different experimental blocks. Half of the lists displayed the blocks in chronological order (1-30), while in the other half, blocks were presented in reverse order (30-1). All participants were presented with both experimental conditions (AV and AV-Blocked). In each list, half of the experimental blocks were presented in the AV condition and half of the experimental blocks were presented in the AV-Blocked condition. The order of presentation of the experimental condition (AV vs AV-Blocked) was counter-balanced between lists. For the test phase, the novel word appeared first in half of the blocks and the target word was presented first in the other half, and the order was randomized for each list. Randomization was created using the list mode on Random.org (2021).

### **2.4. Procedure**

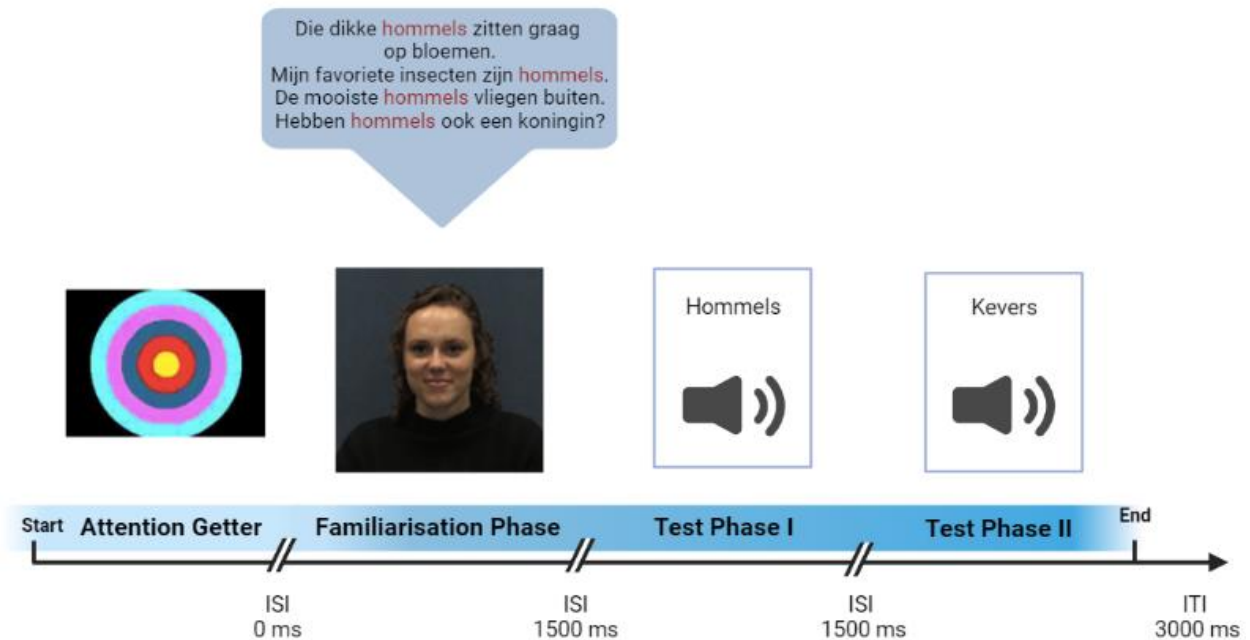
Prior to the lab visit, caregivers were asked to fill in two questionnaires via the online software Castor (2021). One was the Dutch version of the MacArthur-Bates Communicative Development Inventory (N-CDI 1) to assess the infants' receptive and expressive vocabulary size at the time of testing (Zink & Lejaegere, 2002). The other was

a family background questionnaire to assess parental education level, the participants' reading habits and other information about the infants' daily environment.

Each experimental session was run by two experimenters, the lead experimenter and an assistant. Both EEG and eye-tracking data were collected throughout the experiment. The session began with letting the infant sit on a play mat with various toys to become familiar with the new environment. Meanwhile, the experimenter would set up the EEG cap by adding gel to all electrodes while the assistant explained the experimental procedure to the caregiver. The experimenter then fitted the cap on the infant and added gel until the impedance on every electrode was satisfactory. Next, infants and their caregivers were seated in a sound-attenuated experimental booth with a Faraday cage, and the experiment began. The experiment typically started with the eye-tracking calibration unless the baby was too fussy. In these cases, eye-tracking data was not collected and the experiment began immediately.

Each participant was presented with 30 experimental blocks, with 15 blocks in the AV condition and 15 blocks in the AV-Blocked condition. The experiment began with an attention-getter, followed by two ten-second baseline videos of the speaker looking at the infant (one in the AV condition and one in the AV-Blocked condition). The baseline videos were used to get the infant familiar with the visual stimuli and thus contained no speech. Afterwards, the experimental blocks were initiated. Each experimental block contained a familiarisation phase presented in the AV or AV-Blocked condition, followed by the test phase in which two single words were presented auditorily (see Figure 3). The visual speech cue condition of the familiarisation phase was altered after every two blocks (e.g., AV, AV, AV-Blocked, AV-Blocked etc.). The only exceptions were the last two experimental blocks in which the condition was switched after one block to ensure that an equal amount of AV and AV-Blocked videos were presented to each participant. Attention-getters were also displayed before the very first block and then after every two blocks (i.e., whenever the visual speech cue condition switched) to maintain the infants' attention towards the screen. The familiarisation phase began immediately after presentation of the attention getter. The familiarisation phase lasted between 14-21 seconds ( $M = 18.67$  s,  $SD = 1.08$ ). Inter-stimulus intervals between the familiarisation phase and the test phase as well as between the words presented in the test phase were approximately 1500 ms.

Inter-trial intervals between two experimental blocks were 3000 ms. In total, the experiment lasted approximately 15 minutes.



**Figure 3.** Example timeline of an experimental block in the AV condition. Figure shows the presentation of the familiarisation and the test phase with inter-trial intervals (ITI) and inter-stimuli intervals (ISI). Figure created with [BioRender.com](https://www.biorender.com).

The software Presentation (Neurobehavioral Systems, Inc.) was used to run the experiment. Videos were presented at the centre of a 24-inch screen. Auditory stimuli were presented to infants through two loudspeakers in the experimental booth. All stimuli (familiarisation + test phase) were presented at a decibel level of 65-70 dB.

Infants were placed 70 cm away from the screen and sat on their caregiver's laps during the experiment. Caregivers were asked to wear headphones that played distractor music throughout the entire experiment to mask the audio coming from the speakers. Infants were given breadsticks and/or silent toys if they became fussy. The experimenters sat in a neighbouring room and could observe the participants through a video camera. In case the infants became fussy or showed signs of distress, testing was paused and child-friendly videos were played as a distraction. If that did not improve the infants' mood, the

experiment was concluded. Data was collected in the Baby Lab of the Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands.

## **2.5. Recordings**

EEG data was collected from a 32-channel Ag/AgCl ActiCap system. A sampling rate of 500 Hz and an on-line low cut-off filter of 10 Hz and high cut-off of 1000 Hz were used. Electrodes were placed in accordance with the 10/20 system (F7/3/4/8, FC5/1/2/6, T7/T8, C3/4, CP5/1/2/6, TP9/10, P7/3/4/8, Fz, Cz and Pz). FCz was used as the Reference electrode and Fpz as the Ground electrode. Two electrodes were further placed directly on the left and right mastoid bones (TP9L/10L). To capture eye movements, electro-oculogram (EOG) was recorded by placing an electrode on the infants' cheek below their left eye, as well as two electrodes above the eyes (Fp1/2) for vertical EOG and two electrodes to the left and right of the eyes (FT9/10) for horizontal EOG. EEG activity was monitored and recorded on BrainVision Recorder Software (Brain Products GmbH, Germany). Impedance was usually kept below 25 k $\Omega$ .

For some participants, eye-tracking data was also collected throughout the study (data not reported here). Eye-tracking was recorded with EyeLink (Version 5.09). A 16 mm lens and arm mount settings were chosen. Additionally, all experimental sessions were video recorded for later off-line coding.

## **2.6. Analysis**

As previously mentioned, the current sample size of seven infants did not allow us to analyse the influence of visual speech cues on infants' word segmentation. Thus, we here report only the analyses for cortical speech tracking and EEG spectral power which were conducted over the familiarisation phase.

**Exclusion criteria.** Infants were excluded from the analyses if they had fewer than 20% artefact-free trials. Furthermore, participants were excluded if more than four EEG channels had to be removed.

**Looking Times.** Infants' looking times towards the screen were calculated as a measure of the infants' attention towards the stimuli. Because eye-tracking data was not collected for the majority of subjects included in the current sample, infants' looking times were manually coded by the experimenter on ELAN (version 6.2.). Using the video

recordings of the experimental sessions, the experimenter coded for periods in which the infant looked towards the screen as well as periods in which the infant looked away. However, as looking times were manually coded with the recordings of the session and not through the eye-tracking data, it is unclear where exactly infants were looking at when they were looking at the screen. Thus, we cannot make any statements about which facial regions of the speaker the infants were looking at (e.g., the speaker's eyes, mouth or nose) while they were looking at the stimuli.

**Data pre-processing.** EEG pre-processing was conducted with the FieldTrip toolbox (Oostenveld et al., 2011) in MATLAB (Version R2016a; TheMathWorks, Natick, MA, USA). EEG data was cut into 1 second epochs and a high-pass filter at 0.1 Hz as well as a low-pass filter at 30 Hz were applied. Flat channels or channels that displayed large artefacts ( $> \pm 150 \mu\text{V}$  for EEG channels,  $> \pm 250 \mu\text{V}$  for EOG channels) were removed. Independent component analysis (ICA) was performed to remove eye movements and channel artefacts from the data.

**Cortical speech tracking.** Cortical speech tracking was analysed as speech-brain coherence (SBC) over the familiarisation phase. SBC is calculated by first computing the speech envelope in the acoustic signal and then looking at the phase-consistency between the EEG data and the corresponding stimulus envelope of each epoch. For this, the raw EEG data was segmented into four-second epochs with a one-second sliding window throughout the familiarisation materials. Data was low-pass filtered at 45 Hz and high-pass filtered at 1 Hz. For those four-second epochs, the previously identified ICA components were removed and the data was re-referenced to the linked mastoids. The speech envelope was computed using a Hilbert transform with a 2<sup>nd</sup>-order Butterworth filter and added to the EEG data. Then, any remaining trials that displayed large artefacts ( $> \pm 150 \mu\text{V}$ ) were excluded and noisy channels were repaired using spline interpolation and a custom neighbourhood structure. A Fourier transform from 1–10 Hz with a Hanning taper (frequency resolution of .25 Hz) was performed for the speech envelope and the EEG signal. Speech-brain coherence (SBC) was computed as the cross-spectrum between the EEG signal and the speech signal, normalised by their power spectra (Rosenberg et al., 1989).

SBC was first compared between the observed and shuffled data (i.e., a shuffled speech envelope randomly paired to EEG epochs) to investigate whether cortical speech tracking was truly present (i.e. above chance-level) in our data. Secondly, SBC was compared between the AV and AV-Blocked experimental conditions to investigate whether the presence of visual speech cues enhanced infants' cortical speech tracking. To account for differences in the number of trials between the experimental conditions, SBC was normalised for the AV and AV-Blocked data separately as well as the overall data set. This was done with the following formula:

$$\text{Coherence}_{\text{normalised}} = \frac{\text{Coherence}_{\text{observed}} - \text{Coherence}_{\text{shuffled}}}{\text{Coherence}_{\text{observed}} + \text{Coherence}_{\text{shuffled}}}$$

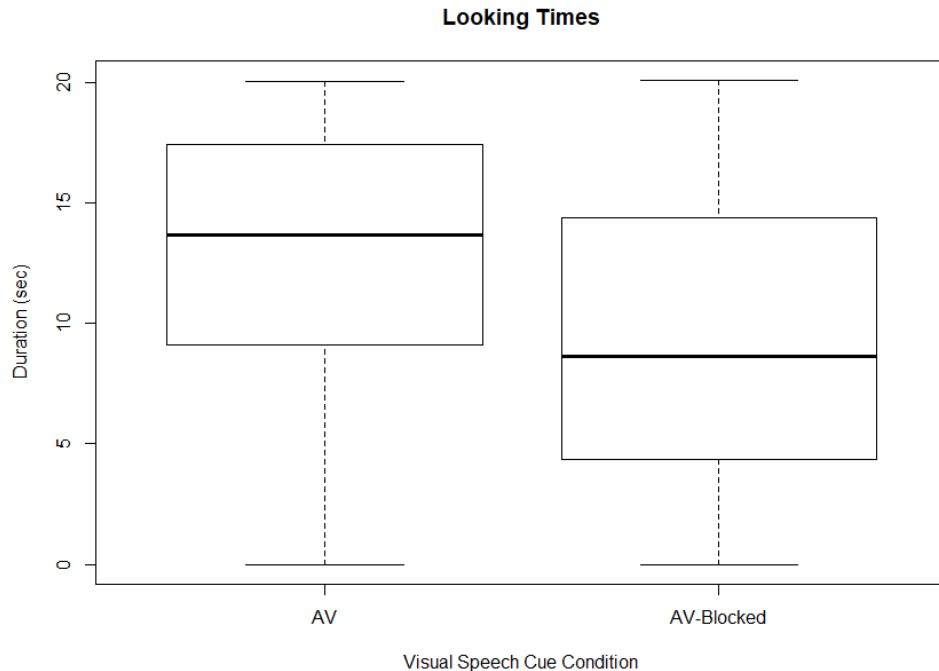
Cluster-based permutation tests (Maris & Oostenveld, 2007) were performed to compare both the shuffled vs observed data as well as the AV vs AV-Blocked data. Cluster-based permutation tests allow us to analyse electrophysiological data over various electrodes and frequencies by conducting paired t-tests while controlling for the multiple comparisons problem (Maris & Oostenveld, 2007; Meyer et al., 2021). T-tests for neighbouring electrodes and frequencies that are found to be significant (i.e., that exceed the pre-defined threshold alpha level of .05) were combined to a cluster. For each cluster, a cluster-level statistic was calculated by summing the individual t-values of the electrodes and frequencies within that cluster. This was then compared to a reference randomisation distribution of the summed cluster t-values. Based on the literature, we decided to focus our analyses on two frequency bands in particular; those were the syllable rate (2.5–3.5 Hz) and the stressed syllable rate (1–1.75 Hz, henceforth referred to as stress rate). For these analyses, cluster-based permutation tests were conducted by averaging over these frequency bands and clustering over electrodes. Furthermore, an exploratory analysis was conducted with a cluster-based permutation test ranging from 1–7 Hz without averaging over frequencies. This was done because this range encompasses all frequency bands in which cortical speech tracking is typically found to be enhanced by visual speech cues in previous studies (Luo et al., 2010; Power et al., 2012; Crosse et al., 2015).

**Spectral Power.** Absolute spectral power was computed with the same Fourier transform over 1-10 Hz (and with a frequency resolution of .25 Hz) as SBC, but the data was averaged over trials. Cluster-based permutation tests were then conducted to compare raw spectral power across various frequencies and channels between the AV and the AV-Blocked condition. These analyses were conducted over the same frequency ranges of interest as SBC: the stress rate (1–1.75 Hz) and the syllable rate (2.5–3.5 Hz) (averaging over those frequencies) and the frequency range from 1–7 Hz (without averaging over frequencies).

### 3. Results

#### 3.1. Behavioural Results

**Looking Times.** Looking times towards the screen were analysed over all trials as a proxy of the infants' attention to the different conditions. A paired t-test revealed that looking times differed significantly between conditions ( $t(6) = 4.62, p < .005$ ). Infants looked significantly longer towards the screen in the AV condition ( $M = 12.72$  seconds,  $SD = 5.5$ ) in which the speaker's full face was visible compared to the AV-Blocked condition ( $M = 8.96$  seconds,  $SD = 5.9$ ) where the speaker's lower face was occluded. This indicated that, overall, infants paid more attention to the stimuli when the speaker's visual speech cues are present.



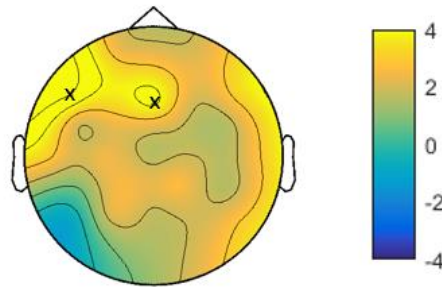
**Figure 4.** Infants' mean looking times towards the screen across visual cue conditions.

#### 3.2. EEG Results

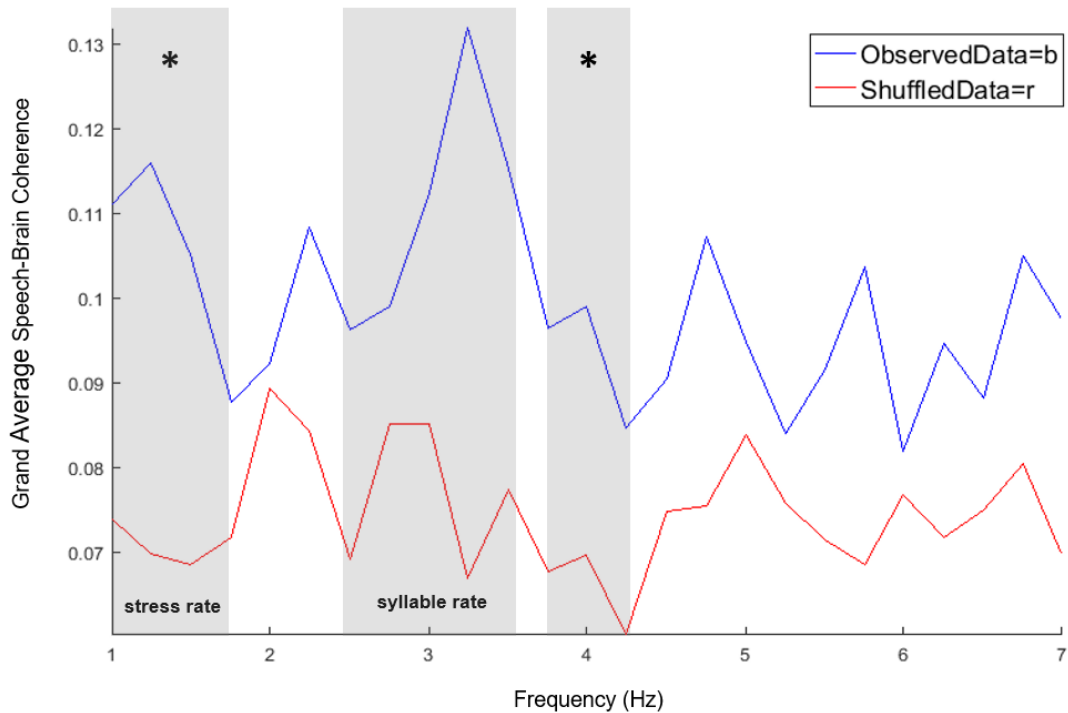
**Shuffled vs Observed Speech-Brain Coherence.** Observed speech-brain coherence (SBC) was compared to shuffled SBC to test whether SBC was truly present in our data. First, cluster-based permutation tests were conducted by averaging over frequencies in our a-priori selected frequency ranges of interest; these were the stress

rate (1–1.75 Hz) and the syllable rate (2.5–3.5 Hz). When looking at the stress rate, significantly higher SBC for the observed data compared to the shuffled data was found in two clusters, specifically around frontal-central electrodes (F7, cluster  $p = .016$ ; Fz, cluster  $p = .029$ , Figure 5). For the syllable rate (averaging over 2.5–3.5 Hz), the cluster-based permutation test revealed no differences between the observed data and the shuffled data, indicating that SBC was not present above-chance at this frequency band.

Stress rate (1 – 1.75 Hz)

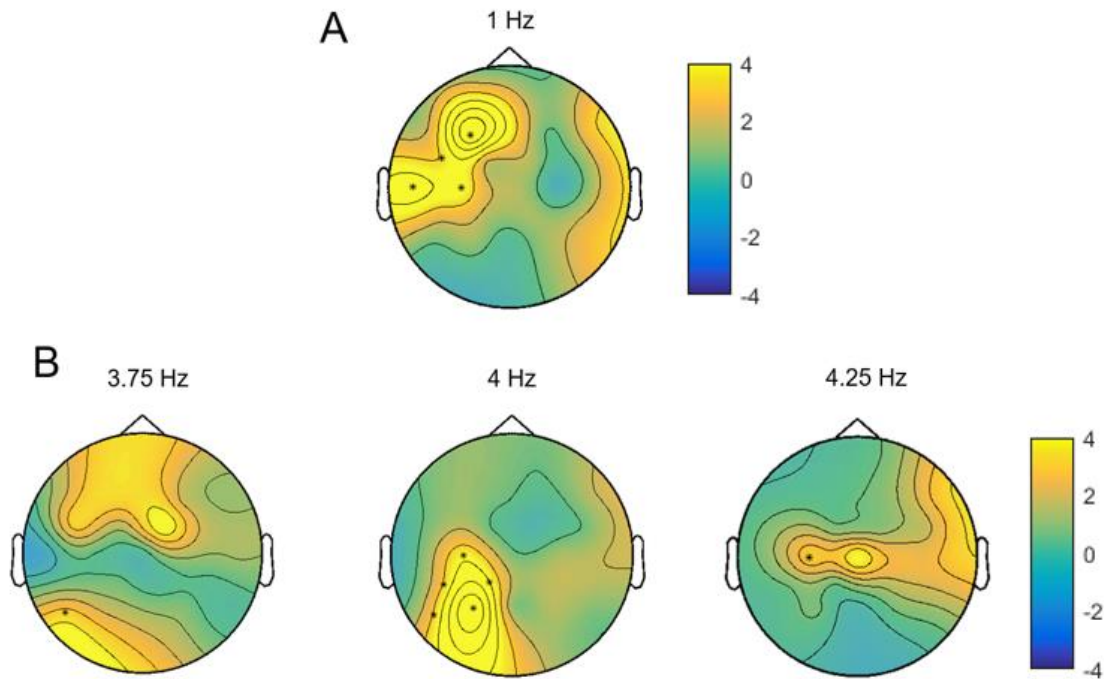


**Figure 5.** Topographic isovoltage map with clusters showing significant SBC in the stress rate. Colours indicate t-statistics of the SBC difference between the shuffled vs observed data. Significant electrodes (F7 and Fz) are marked with x.



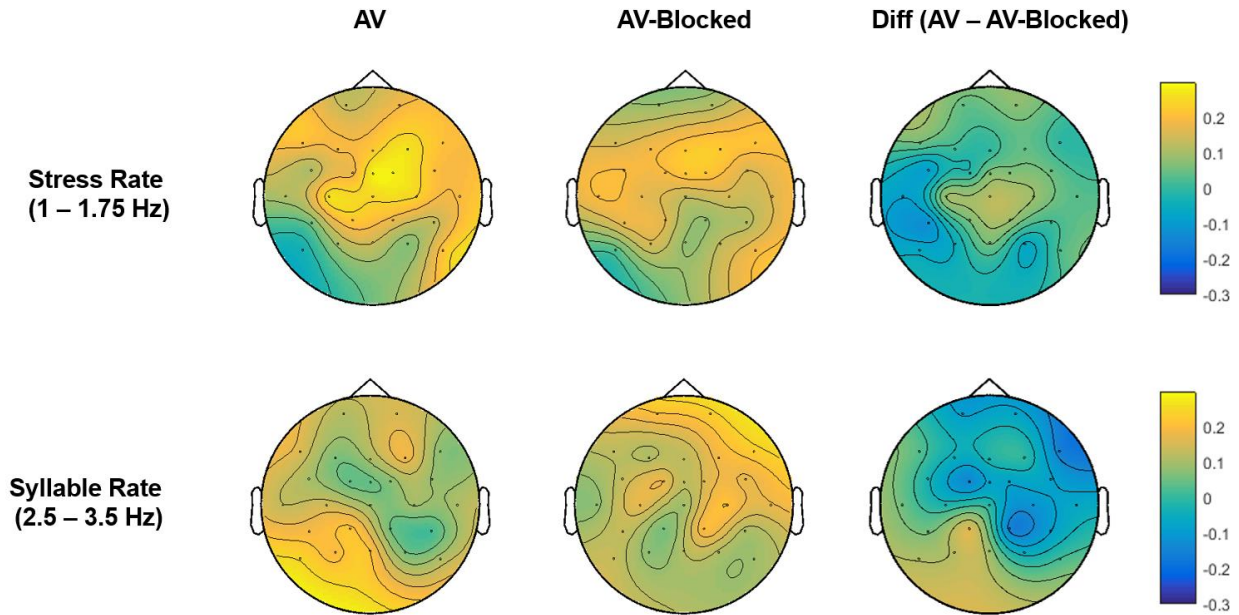
**Figure 6.** Mean Speech-Brain Coherence (SBC) from 1-7 Hz across all EEG channels. SBC between the neural data and the real speech envelope (Observed Data, blue line) and the neural data coupled to a shuffled speech envelope (Shuffled Data, red line) are depicted. The a-priori selected frequency ranges of interest (i.e., the stress rate at 1-1.75 Hz and the syllable rate at 2.5-3.5 Hz) as well as the area in which significant SBC was found are shaded. Ranges with significantly higher observed SBC are marked with a \*.

Moreover, an exploratory analysis was conducted from 1–7 Hz without averaging over frequencies. This revealed significantly higher SBC to the real envelope than to the shuffled envelope (Figure 6). This was found particularly at two frequency bands, one at 1 Hz over left-frontal electrodes (F3, C3, T7, FC5; cluster  $p = .0499$ ), which converges with the stress rate, and the second at 3.75 to 4.25 Hz mostly over left-posterior channels (P7, C3, CP5, CP1, P3;  $p = .009$ ), as depicted in Figure 7. Overall, these results show that SBC was found to be significant at the stress rate and the frequency range between 3.75-4.25 Hz in our data.



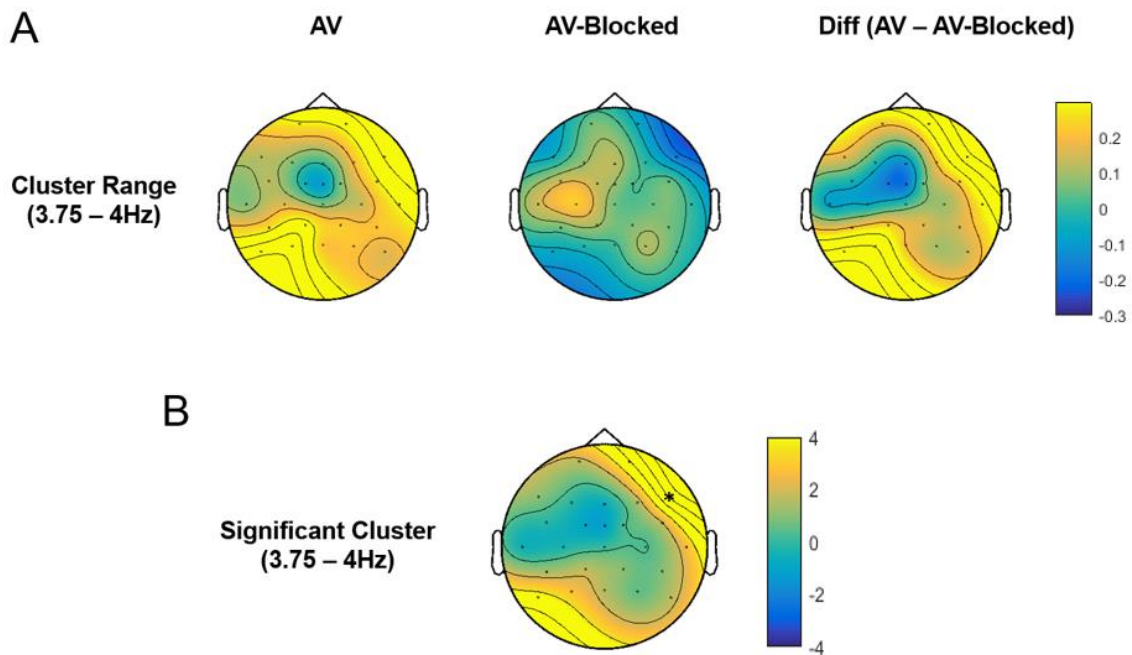
**Figure 7.** Topographic isovoltage maps with significant clusters revealed in the cluster-based permutation test from 1-7 Hz. Colours indicate t-statistics of the speech-brain coherence difference between conditions. Significant electrodes are marked as stars. Results show that SBC of the observed data is significantly higher than SBC of the shuffled data at the frequency bands around 1 Hz **(A)** and 3.75–4.25 Hz **(B)**.

**SBC in AV vs AV-Blocked conditions.** Speech-brain coherence in the two experimental conditions (i.e., AV and AV-Blocked) was also compared using cluster-based permutation tests. When averaging over the frequencies ranges corresponding to the stress (1–1.75 Hz) and syllable rate (2.5–3.5 Hz), no significant differences between the two experimental conditions were found for neither the stress rate nor the syllable rate. This means that SBC did not differ significantly between the conditions at the frequency bands that we predicted. Visual inspection of the topographies (see Figure 8) suggests that SBC might be stronger in the AV condition than the AV-Blocked condition for the stress rate, especially around central channels, but this difference was not found to be statistically significant.



**Figure 8.** Topographic isovoltage maps of raw speech-brain coherence values across the two experimental conditions (AV vs AV-Blocked) as well as the difference between conditions. SBC is shown for the stressed syllable and the syllable rate. None of the differences between the conditions were found to be significant for those frequency bands.

Furthermore, a cluster-based permutation test was performed over all frequencies between 1–7 Hz (without averaging over frequencies). This revealed that SBC was significantly higher in the AV condition than the AV-Blocked condition in the frequency band between 3.75–4 Hz, (channel F8, cluster  $p = .026$ ; see Figure 9B). Thus, while SBC was found to be significantly higher in the AV condition (i.e. the condition in which visual speech cues were visible to the infants) than in the AV-Blocked condition as expected, this effect was found in a different frequency range than we predicted.

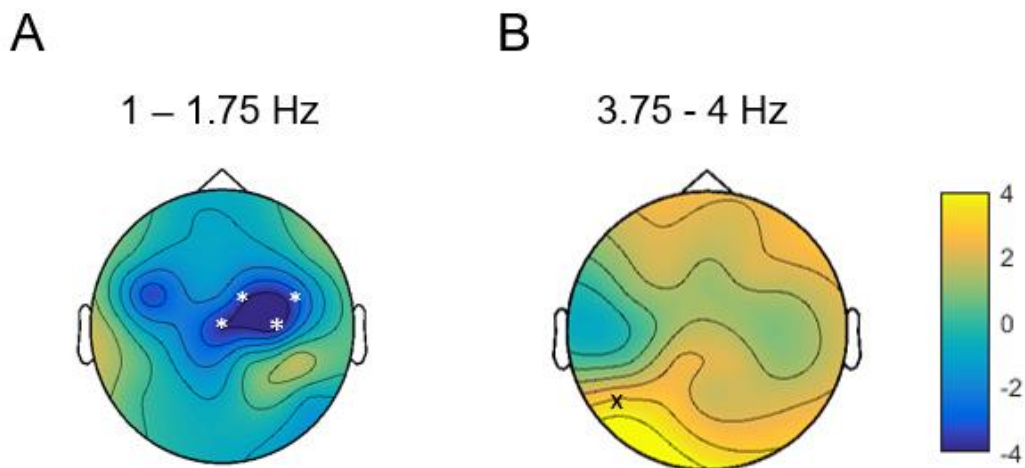


**Figure 9. A.** Topographic isovoltage maps of raw speech-brain coherence values across the two experimental conditions (AV vs AV-Blocked) as well as the difference between conditions from 3.75-4 Hz. **B.** Topographic isovoltage map showing the significant positive cluster found around channel F8 (marked as star), indicating higher SBC in the AV condition than the AV-Blocked condition. Colours indicate t-values of the speech-brain coherence (SBC) difference between conditions.

To check whether SBC was indeed more present in the AV than the AV-Blocked condition at the frequency range in which the significant cluster was found (3.75–4 Hz), a post-hoc analysis was conducted comparing observed vs shuffled SBC in both the AV and AV-Blocked conditions separately. First, observed vs shuffled SBC was compared for EEG data recorded only in the AV condition by averaging over the frequency range from 3.75–4 Hz. Observed SBC was found to be significantly higher than shuffled SBC at 3.75–4 Hz, specifically at channels CP1 and P3 (cluster  $p = .014$ ) and channel P7 (cluster  $p = .042$ ). However, when looking only at the EEG data recorded over the AV-Blocked condition, no difference in SBC was found between the observed and shuffled data between 3.75–4 Hz. This suggests that significant SBC is mostly present in the AV condition compared to the AV-Blocked condition, converging with our results that SBC is

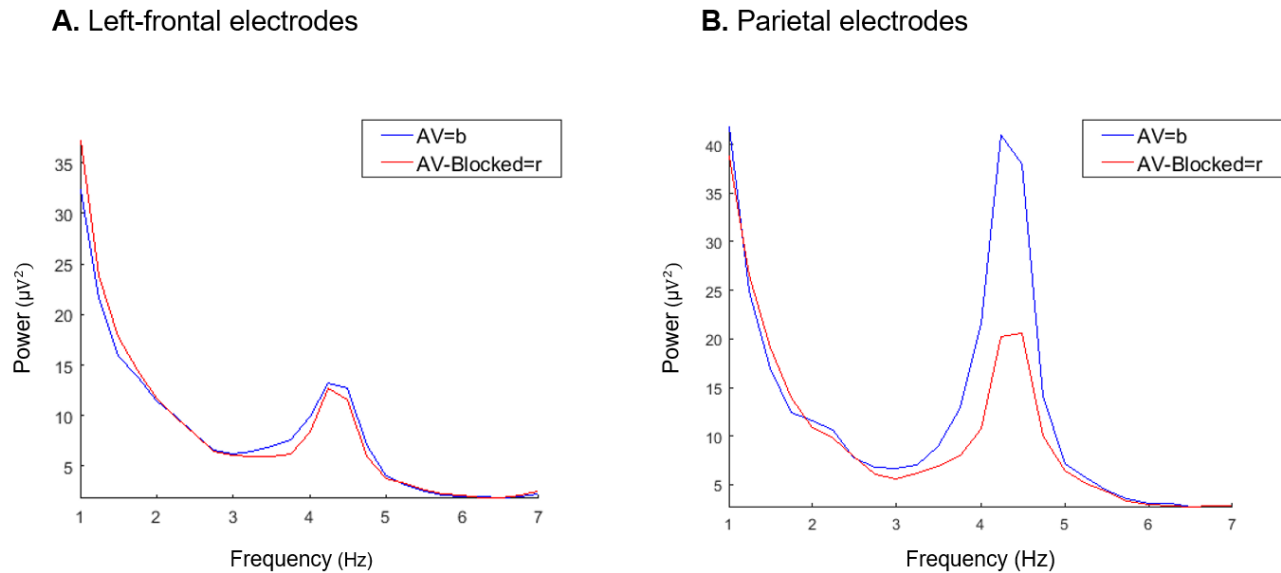
significantly stronger when infants can see the speaker's visual speech cues at this frequency band.

**Spectral power.** Absolute spectral power was also compared between the AV and the AV-Blocked condition. Running a cluster-based permutation test across all frequencies from 1–7 Hz without averaging over frequencies or channels, no significant differences in spectral power between the conditions were found. However, cluster-based permutation tests were also conducted in our frequency ranges of interest; these were again the stress rate (1–1.75 Hz) and the syllable rate (2.5–3.5 Hz), as well as at the frequency range in which the significant clusters were found in the speech-brain coherence analysis (3.75–4 Hz). While there were no differences in the power spectra of the AV and AV-Blocked conditions at the syllable rate, the tests revealed that power was significantly higher in the AV-Blocked condition than in the AV condition for the stress rate over right-central electrodes (cluster  $p = .001$ ; Figure 10A). Lastly, averaging over the frequency band from 3.75–4 Hz, significantly higher raw spectral power was found in the AV condition than the AV-Blocked condition, particularly over channel P7 (cluster  $p = .034$ ; Figure 10B).



**Figure 10.** Topographic isovoltage maps for the spectral power analysis. Colours indicate t-statistics of the SBC difference between conditions found in the cluster-based permutation test. **A.** Spectral power was significantly lower for the AV condition in the stress rate (left), particularly around channels FC2, C4, Cz, FC6 (marked as stars). **B.** Conversely, spectral power was found to be significantly higher for the AV condition at the range between 3.75–4 Hz, specifically over channel P7 (marked as x).

Visual inspection of raw spectral power indicated that this difference can be seen especially in parietal regions (see Figure 11B), while the power spectra are equally strong in both conditions over left-frontal areas (Figure 11A). Interestingly, a peak in the power spectra at frequencies around 3.5–5 Hz is found for both conditions, and over both left-frontal as well as parietal areas.



**Figure 11.** Absolute spectral power across all frequencies ranging from 1–7 Hz in the AV condition compared to the AV-Blocked condition. **A.** Power over left frontal areas (i.e., channels F7, F3, FC5). **B.** Power over parietal areas (i.e., channels P3, Pz, P4).

#### 4. Discussion

The current study investigated whether visual speech cues enhance 10-month-old infants' cortical tracking of speech. Infants' neural activity was measured with EEG while they saw videos of a female speaker presented in two different experimental conditions: one in which the speaker's audio-visual speech cues were fully visible (AV condition) and another in which the speaker's lower face (including her lips, mouth and jaw) was blocked (AV-Blocked condition). We predicted that cortical speech tracking would be higher when visual speech cues were visible, particularly at the stress rate (1–1.75 Hz) and the syllable rate (2.5–3.5 Hz).

First, we found significantly higher speech-brain coherence (SBC) to the observed data compared to the shuffled data at the stress rate (1–1.75 Hz) as well as the frequency range between 3.75 and 4.25 Hz. This suggests that SBC was truly present in our data, although not in all frequencies. Following this, we compared SBC in the AV and AV-Blocked conditions. The results did not reveal any differences in SBC between the conditions in neither the stressed syllable nor the syllable rate, contrary to our predictions. However, the results showed significantly higher SBC in the AV condition at the frequency range between 3.75–4 Hz. This means that infants' cortical speech tracking was significantly higher when they could see the speaker's visual speech cues, as predicted by our hypothesis, but at a different frequency range than expected. Instead, this effect was found at a frequency range slightly above the syllable rate, corresponding to the theta band in infants (commonly defined as 3–6 Hz (Orehova et al., 2006; Meyer et al., 2019)).

Furthermore, we investigated whether absolute spectral power of the EEG signal differed between conditions. While spectral power was significantly higher in the AV-Blocked condition for the stressed syllable rate, power was significantly higher in the AV condition at the frequency range between 3.75–4 Hz (i.e., the same range in which SBC was stronger for the AV condition).

Lastly, infants' looking times towards the screen were compared between the two conditions as a proxy for the infants' attention towards the stimuli. The results showed that infants looked significantly longer towards the screen when the speaker's full face was visible compared to the condition in which the speaker's articulatory cues were blocked.

Overall, our results suggest that visual speech cues (i.e., the rhythmic movements of the mouth, lips and jaw) may increase infants' attention to the stimuli as well as enhance speech-brain coherence and EEG spectral power at the theta band, particularly around 3.75–4 Hz.

#### **4.1. Visual speech cues enhance speech-brain coherence at theta band**

Our results indicate that infants' cortical speech tracking was enhanced when visual speech cues were visible at the 3.75–4 Hz frequency range, but not at the ranges corresponding to the stress and syllable rate. Strikingly, this frequency range converges well with the findings reported in adult studies that visual speech enhances cortical speech tracking at frequencies between 2–7 Hz (Luo et al., 2010; Crosse et al., 2015, Aller et al., 2021). However, our results converge with the adult findings in a different way than expected. In adult studies, it is often claimed that visual speech cues enhance entrainment to the speech envelope at frequencies around 2–7 Hz (i.e. delta-theta frequencies) as those are the frequencies in which stressed and unstressed syllables appear in ADS and thus in which the mouth movements are locked to the speech envelope. However, in our study, which used IDS, we computed that stressed syllables and syllables appeared at lower rates, particularly at 1–1.75 Hz and 2.5–3.5 Hz. Nonetheless, cortical speech tracking did not differ between the conditions at those rates. Instead, enhanced cortical speech tracking was again found around the theta range similarly to the findings reported in adults.

One possible explanation for this effect appearing at around 4 Hz is that the visual speech envelope (i.e., the lip aperture envelope) may peak around 4 Hz in our stimuli, even if this does not correspond to the stress or syllable rate. In this study, we did not compute the visual speech envelope to investigate this possibility. However, in studies that have analysed both the visual and acoustic speech signal, the visual speech envelope is typically found to peak at slightly lower frequencies than the auditory speech envelope (Park et al., 2016; Aller et al., 2021). Thus, if our results would be due to entrainment towards the visual speech envelope, we would expect to find this effect at slightly lower frequencies than the syllable rate, and not in higher frequencies. Therefore, it may not be the properties of the (audio-visual) speech envelope *per se* that lead to higher speech-

brain coherence at the theta range, but it may be something about these neural oscillatory bands in particular.

More specifically, one possible explanation for these unexpected results is that oscillatory activity at the theta range may be responsible for audio-visual integration during speech processing, regardless of what the speech input is. This would suggest that adults show higher cortical speech tracking for audio-visual speech in the frequencies between 2–7 Hz not because that is the syllable rate in ADS, but because neural activity at the theta band drives successful integration of the two modality streams (audio and visual) in those conditions. Support for this theory comes from a recent paper that suggests that the visual and auditory cortices may communicate with each other through theta-band synchronisation during audio-visual speech perception (Biau et al., 2021). In particular, the paper proposes that the visual speech signal (which precedes the auditory signal in natural speech) induces increased firing of theta-band oscillations in the visual cortex, which in turn enhance the excitability of the auditory cortex. Thus, these theta oscillations prepare the auditory cortex for the incoming speech signal, allowing it to arrive at an optimal time point of high neuronal excitability (Biau et al., 2021). Consequently, this results in oscillatory activity being optimally phase-locked to the speech envelope which could lead to better cortical speech tracking in conditions where visual speech cues are present, as found here. Therefore, it is possible that communication between the visual and auditory cortices through theta-band activity may be used as a predictive mechanism during audio-visual speech processing, enhancing speech-brain coherence.

Furthermore, our results converge with the study by Power et al. (2012) on 13-year-old children's entrainment to single syllables presented in audio-visual, audio-only and visual-only conditions. Their results showed that 13-year-olds' entrainment to the single syllables was significantly higher in the audio-visual condition in the theta band (4 Hz), but not in the delta band (2 Hz). Crucially, syllables were presented at a standard frequency of 2 Hz (including the visual articulation of the syllable) in this study. If cortical speech tracking is enhanced by visual speech cues at the rate in which syllables appear and thus the mouth movements are locked to the speech stimuli (as claimed by the adult studies), one would predict the opposite effect, namely that entrainment would be enhanced in the delta band instead of the theta band. The authors of this paper do not offer an explanation

for this mismatched finding, but it is possible that their results are due to audio-visual integration being modulated by theta-band oscillations in their subjects as well.

However, this interpretation should be handled with caution as we currently only find significantly higher SBC for the AV condition over one frontal channel (F8), and not over temporal or posterior regions. Visual inspection of the clusters and topographic maps reveal that SBC also appears to be stronger over posterior areas, but this was not found to be significant. With more data, it will hopefully be possible to investigate this theory further. If theta-band activity between the visual and auditory cortices is indeed used for audio-visual integration, we would expect to see significant clusters further emerge over temporal and posterior regions with more statistical power.

#### **4.2. Visual speech cues modulate spectral power at the delta and theta range**

Our results demonstrated that spectral power was also increased in the theta-band when audio-visual speech cues were fully visible. In particular, enhanced spectral power for the AV condition was found over posterior areas around channel P7. This finding provides further support for the possibility that audio-visual integration may be driven by theta-band oscillations during multi-modal speech processing. The increased spectral power over posterior regions found here could be due to increased firing of oscillatory activity in the theta range over the visual cortex which is then used for communication with the auditory cortex, as proposed by Biau et al. (2021).

Interestingly, higher spectral power was found in the AV-Blocked condition compared to the AV condition at the stress rate (1–1.75 Hz). One possible explanation for this finding is that having articulatory cues occluded may bias the infants' attention towards salient properties of the speech signal, such as prosodic cues. One of the main prosodic cues that infants at 10 months are sensitive to is metrical stress (Jusczyk et al., 1993). Thus, stressed syllables may become especially salient in the absence of audio-visual speech input, enhancing the infants' neural firing at those rates in particular.

#### **4.3. Methodological considerations and possible confounds**

**Sample size.** The results presented here are based on the data collected from seven infants. While our preliminary results already indicate significant SBC differences

between the conditions, it is entirely possible that the results will change with a larger data set and increased statistical power. For instance, it is possible that with more statistical power, differences in SBC between the conditions will also arise at other frequency bands, such as the stress and syllable rate. It is also possible that the effects reported here are due to noise in the data, and may disappear again with more statistical power. Furthermore, with a larger sample size, it is also possible that significant clusters may emerge over different brain areas than the ones reported here. This would allow us to further test the idea that theta-band oscillations may drive communication between the visual and auditory cortices during audio-visual speech processing. Data for this study will continue to be collected for the coming months, with plans to collect data from 40 infants in total. With the full sample, we will hopefully be able to determine whether infants' SBC is truly enhanced by visual speech cues and in which frequency ranges and brain regions these effects exactly occur.

**Higher attention to AV speech.** In our current dataset, infants looked significantly longer towards the screen when the speaker's full face was visible compared to the videos in which the speaker's visual speech cues were occluded. This indicates that infants may be more engaged with the stimuli and thus pay more attention to the speech envelope in the AV condition. This difference in attention may be a possible confound for the findings reported here. Both cortical speech tracking and spectral power have been found to be modulated by attention. For instance, various studies have demonstrated that adults' neural entrainment to attended speech is higher than entrainment to unattended speech (Ding & Simon, 2012; Zion-Golumbic et al., 2013; Rimmele et al., 2015). Similarly, spectral power, especially in the theta band, has also been found to increase with attention and emotional arousal in children (Maulsby, 1971; Xie et al., 2018; Meyer et al., 2019). Therefore, it is unclear whether our results are due to visual speech cues directly facilitating cortical speech tracking, or because seeing the speaker's mouth movements enhances infants' arousal towards the stimuli, and it is this higher attention that modulates their ability to process the speech envelope.

As our current sample size was limited, we decided to include all available trials into our analysis. Therefore, we could not control for attentional differences between the conditions. However, with a larger data set, there are multiple ways to control for

attentional differences between the conditions. For instance, we could only include trials in which the infant was looking at the stimuli for a specific amount of time (e.g., 50% of the trial). Another possibility would be to add attention as a co-variate in our model. These options would allow us to tease apart whether visual speech cues directly influence SBC and spectral power or whether the effects reported here are driven by differences in attention.

Nonetheless, the fact that infants look less towards the screen when visual speech cues are occluded is an incredibly interesting finding in its own right. In fact, this difference in attention may be part of the mechanism through which lower SBC was found in the AV-Blocked condition. As mentioned in the introduction, infants between six and twelve months of age show preferential looking towards a speaker's mouth and lips (Tenenbaum et al., 2013; Lewkowicz & Hansen-Tift, 2012). Thus, having these facial regions occluded during face-to-face communication may make infants less interested in the interaction and lead them to shift their attention away from the speaker and speech input. This in turn could have negative effects on their ability to process and track the speech envelope.

**Spectral power differences.** In this study, enhanced spectral power at the theta range was found for the AV condition. However, it is possible that our main result, namely that SBC is higher for the AV condition at 3.75–4 Hz, is directly caused by those differences in spectral power rather than by SBC being modulated by visual speech cues. More specifically, because coherence was computed with the power spectra of the EEG data, significant differences in spectral power may have led to the differences we find in SBC. This is especially the case if the EEG data used to compute coherence has different signal-to-noise ratios across different electrodes (Bastos & Schoffelen, 2016) which is certainly a possibility with noisy infant data. The fact that EEG spectral power can influence coherence is the case for any study and should always be addressed when interpreting results. However, as more data will be collected for the current study, this possibility should be minimised with a larger sample by increasing statistical power and decreasing noise.

#### **4.4. Implications and future directions**

Our current findings provide further evidence for infants' sensitivity to multi-modal cues, particularly facial cues, from early on in life. These results have important implications especially for infants growing up during the COVID-19 pandemic where interactions with face masks (which occlude the same parts of the speaker's face as in our materials, except for the larynx) are increasingly common. For instance, our results indicate that infants may pay less attention to the speech input when a speaker's articulatory cues are blocked, which could potentially have negative effects on their language development. Future studies should further investigate the effects of face masks on infants' language acquisition, such as their vocabulary development. Moreover, in our study, the auditory materials were identical between the two experimental conditions. However, in real life, face masks do not only block out access to visual speech cues, but they also impair the acoustic quality of the speech signal (Nguyen et al., 2021). Therefore, the effects of face masks on speech intelligibility and comprehension should be further explored in future studies, for both adults and infants.

Our second main finding is the possible role of oscillatory activity at the theta range during audio-visual speech integration. Our results indicate that, contrary to the interpretations of many adult studies, it may not be the correspondence between the mouth movements and the speech envelope at the syllable rate that drives higher entrainment between 2–7 Hz. Instead, oscillatory activity at those frequency bands may potentially drive successful audio-visual integration in multi-modal settings by enhancing communication between the visual and auditory cortices. Thus, it would be interesting to test adults' cortical tracking of slowed (or infant-directed) audio-visual speech where the syllable rate does not correspond to the theta band (as is the case for our stimuli). If adults would then show neural tracking of the speech envelope at lower frequencies (e.g. 1–3.5 Hz), then this may be due to their tracking of the linguistic units present in the stimuli. However, if adults would still show higher SBC for audio-visual speech around the theta band (4-8 Hz), this would support the idea that those oscillations may arise more strongly during audio-visual speech integration. Future research should investigate ways to disentangle those two possibilities, and explore how tracking to slowed audio-visual speech may occur across different regions, such as the visual and auditory cortices.

## 5. Conclusion

The current study investigated whether infants' neural tracking of the speech envelope is modulated by the presence of audio-visual speech cues. When visual speech cues were present, infants' cortical speech tracking and EEG spectral power were enhanced at the frequency band around 3.75–4 Hz. These findings introduce the possibility that theta-band oscillatory activity may facilitate cross-modal integration of the visual speech signal (i.e., the articulatory cues of the mouth and lips) and the amplitude modulations of the speech envelope. Conversely, occluding visual speech cues led to lower speech-brain coherence and spectral power at that theta band, but led to higher EEG power at the stressed syllable rate (1–1.75 Hz). Thus, while audio-visual speech cues may elicit neural firing at the theta band, having these cues occluded may bias the infants' attention towards salient prosodic cues in the speech signal, such as stressed syllables. Lastly, we also found that visual speech cues significantly increase infants' looking times towards the screen compared to the condition in which these cues are blocked. This may reflect higher attention to the speech stimuli and therefore result in better speech processing in the AV condition. While we here report preliminary findings based on seven subjects, this thesis is part of a larger on-going study. With a larger sample size, we will be able to disentangle whether the results found here are truly due to the visual speech cues modulating infants' theta-band power and cortical speech tracking or whether these effects are due to attentional differences. Regardless, current findings already provide strong evidence for infants' sensitivity to multi-modal cues and highlight the importance of visual speech cues on infants' neural processing of speech.

## 6. References

- Aller, M., Okland, H. S., MacGregor, L. J., Blank, H., & Davis, M. H. (2021). Visual speech is processed differently in auditory and visual cortex: evidence from MEG and partial coherence analysis. *bioRxiv*. DOI: <https://doi.org/10.1101/2021.12.18.472955>
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database. Linguistic data consortium*. Philadelphia, PA: University of Pennsylvania. DOI: <https://doi.org/10.35111/g6s-gm48>
- Barajas, M. C. O., Guevara, R., & Gervain, J. (2021). The origins and development of speech envelope tracking during the first months of life. *Developmental cognitive neuroscience*, 48, 100915. DOI: <https://doi.org/10.1016/j.dcn.2021.100915>
- Bastos, A. M., & Schoffelen, J. M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9, 175. DOI: <https://doi.org/10.3389/fnsys.2015.00175>
- Biau, E., Wang, D., Park, H., Jensen, O., & Hanslmayr, S. (2021). Auditory detection is modulated by theta phase of silent lip movements. *Current Research in Neurobiology*, 2, 100014. DOI: <https://doi.org/10.1016/j.crneur.2021.100014>
- Biorender (2022). Available at: <https://biorender.com/>.
- Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer*. Retrieved from [www.praat.org](http://www.praat.org).
- Brooks, P. J., & Kempe, V. (2012). *Language development*. John Wiley & Sons.
- Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 45(4), 204-220. DOI: [10.1002/dev.20032](https://doi.org/10.1002/dev.20032)
- Castor EDC. (2019). Castor Electronic Data Capture. Available at: <https://castoredc.com>.
- Chandrasekaran, C., Trubanova, A., Stillitano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology*, 5(7), e1000436. DOI: <https://doi.org/10.1371/journal.pcbi.1000436>

- 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. DOI: <https://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. DOI: <https://doi.org/10.1523/JNEUROSCI.1396-16.2016>
- Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19), 2457-2465. DOI: <https://doi.org/10.1016/j.cub.2015.08.030>
- Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29), 11854-11859. DOI: <https://doi.org/10.1073/pnas.1205381109>
- 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. DOI: <https://doi.org/10.1038/nn.4186>
- ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from <https://archive.mpi.nl/tla/elan>.
- Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of speech and hearing disorders*, 40(4), 481-492. DOI: <https://doi.org/10.1044/jshd.4004.481>
- Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4), 511. DOI: <https://doi.org/10.1038/nn.3063>
- Haegens, S. (2020). Entrainment revisited: a commentary on Meyer, Sun, and Martin (2020). *Language, cognition and neuroscience*, 35(9), 1119-1123. DOI: <https://doi.org/10.1080/23273798.2020.1758335>
- 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. DOI: <http://www.jstor.org/stable/3696454>

- Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child development*, 64(3), 675-687. DOI: <https://doi.org/1131210>
- Kalashnikova, M., Peter, V., Di 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. DOI: <https://doi.org/10.1038/s41598-018-32150-6>.
- Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, 18(7), 1560-1574. DOI: <https://doi.org/10.1093/cercor/bhm187>
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, 9(2), F13-F21. DOI: <https://doi.org/10.1111/j.1467-7687.2006.00468.x>
- Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, 218(4577), 1138-1141. DOI: [10.1126/science.1175626](https://doi.org/10.1126/science.1175626)
- 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, USA*, 105(32), 11442–11445. DOI: <https://doi.org/10.1073/pnas.0804275105>
- Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, 53(2), 279-292. DOI: <https://doi.org/10.1016/j.neuron.2006.12.011>
- 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. DOI: <https://doi.org/10.1073/pnas.1114783109>
- Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using Low-Frequency Neuronal Phase Modulation. *PLoS Biology*, 8(8), e1000445. DOI: <https://doi.org/10.1371/journal.pbio.1000445>

- Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Frontiers in psychology*, 7, 52. DOI: <https://doi.org/10.3389/fpsyg.2016.00052>
- Maulsby, R. L. (1971). An illustration of emotionally evoked theta rhythm in infancy: Hedonic hypersynchrony. *Electroencephalography and Clinical Neurophysiology*, 31(2), 157-165. DOI: [https://doi.org/10.1016/0013-4694\(71\)90186-6](https://doi.org/10.1016/0013-4694(71)90186-6)
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, 164(1), 177-190. DOI: <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748. DOI: <https://doi.org/10.1038/264746a0>
- Mercure, E., Kushnerenko, E., Goldberg, L., Bowden-Howl, H., Coulson, K., Johnson, M. H., & MacSweeney, M. (2019). Language experience influences audiovisual speech integration in unimodal and bimodal bilingual infants. *Developmental science*, 22(1), e12701. DOI: <https://doi.org/10.1111/desc.12701>
- Meyer, M., Endedijk, H. M., Van Ede, F., & Hunnius, S. (2019). Theta oscillations in 4-year-olds are sensitive to task engagement and task demands. *Scientific reports*, 9(1), 1-11. DOI: <https://doi.org/10.1038/s41598-019-42615-x>
- Meyer, M., Lamers, D., Kayhan, E., Hunnius, S., & Oostenveld, R. (2021). Enhancing reproducibility in developmental EEG research: BIDS, cluster-based permutation tests, and effect sizes. *Developmental Cognitive Neuroscience*, 52, 101036. DOI: <https://doi.org/10.1016/j.dcn.2021.101036>
- Micheli, C., Schepers, I. M., Ozker, M., Yoshor, D., Beauchamp, M. S., & Rieger, J. W. (2020). Electroencephalography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. *European Journal of Neuroscience*, 51(5), 1364–1376. DOI: <https://doi.org/10.1111/ejn.13992>
- 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. DOI: <https://doi.org/10.1080/01690965.2013.791703>

- Myers, B. R., Lense, M. D., & Gordon, R. L. (2019). Pushing the envelope: Developments in neural entrainment to speech and the biological underpinnings of prosody perception. *Brain sciences*, 9(3), 70. DOI: <https://doi.org/10.3390/brainsci9030070>
- Nguyen, D. D., McCabe, P., Thomas, D., Purcell, A., Doble, M., Novakovic, D., ... & Madill, C. (2021). Acoustic voice characteristics with and without wearing a facemask. *Scientific reports*, 11(1), 1-11. DOI: <https://doi.org/10.1038/s41598-021-85130-8>
- 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. DOI: <https://doi.org/10.1155/2011/156869>
- Orekhova, E. V., Stroganova, T. A., Posikera, I. N., & Elam, M. (2006). EEG theta rhythm in infants and preschool children. *Clinical neurophysiology*, 117(5), 1047-1062. DOI: <https://doi.org/10.1016/j.clinph.2005.12.027>
- Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife*, 5, e14521. DOI: <https://doi.org/10.7554/eLife.14521>
- Patterson, M.L., & Werker, J.F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, 22(2), 237–247. DOI: [https://doi.org/10.1016/S0163-6383\(99\)00003-X](https://doi.org/10.1016/S0163-6383(99)00003-X)
- Patterson, M.L., & Werker, J.F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6(2), 191–196. DOI: <https://doi.org/10.1111/1467-7687.00271>
- Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*, 3, 320. DOI: <https://doi.org/10.3389/fpsyg.2012.00320>
- Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, 68, 169–181. DOI: <https://doi.org/10.1016/j.cortex.2015.03.006>

- Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Current opinion in neurobiology*, 28, 142-149. DOI: <https://doi.org/10.1016/j.conb.2014.07.005>
- Power, A. J., Mead, N., Barnes, L., & Goswami, U. (2012). Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. *Frontiers in Psychology*, 3, 216. DOI: <https://doi.org/10.3389/fpsyg.2012.00216>
- Random.org. (September, 2021). Available at: <https://www.random.org/lists/>.
- Rimmele, J. M., Golumbic, E. Z., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, 68, 144-154. DOI: <https://doi.org/10.1016/j.cortex.2014.12.014>
- Rosenberg, J., Amjad, A., Breeze, P., Brillinger, D., & Halliday, D. (1989). The fourier approach to the identification of functional coupling between neuronal spike trains. *Progress in biophysics and molecular biology*, 53 (1), 1–31. DOI: [https://doi.org/10.1016/0079-6107\(89\)90004-](https://doi.org/10.1016/0079-6107(89)90004-)
- Snijders, T. M. (2020). Individual variability in infants' cortical tracking of speech rhythm relates to their word segmentation performance. Poster presented at *the Twelfth Annual (Virtual) Meeting of the Society for the Neurobiology of Language (SNL 2020)*.
- 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. DOI: <https://doi.org/10.1121/1.1907309>
- Tan, S. J., Crosse, M. J., Di Liberto, G. M., & Burnham, D. (2019). Four-Year-Olds' Cortical Tracking to Continuous Auditory-Visual Speech. In *Proc. The 15th International Conference on Auditory-Visual Speech Processing*, 53-56. DOI: [10.21437/AVSP.2019-11](https://doi.org/10.21437/AVSP.2019-11)
- Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108(3), 850-855. DOI: <https://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. DOI: <https://doi.org/10.1111/j.1532-7078.2012.00135.x>
- Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. 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. DOI: [10.1017/S0305000914000725](https://doi.org/10.1017/S0305000914000725)
- Xie, W., Mallin, B. M., & Richards, J. E. (2018). Development of infant sustained attention and its relation to EEG oscillations: an EEG and cortical source analysis study. *Developmental Science*, 21(3), e12562. DOI: <https://doi.org/10.1111/desc.12562>
- 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. DOI: <https://doi.org/10.1111/j.1467-7687.2009.00833.x>
- Zheng, Y., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech?. *Applied Psycholinguistics*, 40(1), 93-109. DOI: [10.1017/S0142716418000462](https://doi.org/10.1017/S0142716418000462)
- Zink, I., & Lejaegere, M. (2002). *N-CDIs: Lijsten voor communicatieve ontwikkeling*. Leeuven/Leusden: Acco.
- Zion Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5), 980-991. DOI: <https://doi.org/10.1016/j.neuron.2012.12.037>
- Zoefel, B. (2021). Visual speech cues recruit neural oscillations to optimize auditory perception: Ways forward for research on human communication. *Current Research in Neurobiology*. DOI: <https://doi.org/10.1016/j.crneur.2021.100015>

## 7. Supplementary Materials

Praat was used to determine the duration, pitch (measured as  $f_0$ ) and intensity of all 60 test words as well as the 240 sentences in the familiarisation phase (120 sentences in Version A, 120 sentences in Version B). The mean and SD values for the test words can be found in Table S1.

**Table S1.** Acoustic properties of the test phase stimuli. Mean and SD were computed over all items.

	Test words	
	Mean	SD
Duration (sec)	0.92	0.13
Pitch (Hz)	264.98	63.37
Intensity (dB)	66.72	2.58

For the familiarisation phase, means between the A and B versions were compared using independent t-tests for the data that was normally distributed and Wilcoxon rank sum tests for the data that did not fulfil the normality assumption.

Mean duration did not differ between the two versions for any of the tiers (sentence:  $t(237) = -0.29, p = .77$ ; word:  $W = 320267, p = .85$ ; syllable:  $W = 683106, p = .85$ ).

Equally, mean pitch did also not differ between version A and B for any tier (sentence:  $W = 3136, p = .08$ ; word:  $W = 241108, p = .86$ ; syllables:  $W = 590144, p = .31$ ).

Lastly, mean intensity did not also not differ between the two versions for any of the tiers (sentence:  $t(238) = -0.08, p = .94$ , words:  $W = 311628, p = .52$ ; syllables:  $W = 684806, p = .98$ ). Thus, none of the acoustic measurements analysed differed significantly between the two versions for any tier.

**Table S2.** Acoustic properties of the familiarisation phase stimuli. Mean and SD were computed separately over all items in Version A and Version B and for each separate tier (sentences, words and syllables tiers).

		<b>Familiarisation Phase</b>			
		<b>Version A</b>		<b>Version B</b>	
		<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
<b>Duration (sec)</b>	Sentences	3.19	0.53	3.21	0.49
	Words	0.48	0.26	0.48	0.26
	Syllables	0.33	0.16	0.33	0.16
<b>Pitch (Hz)</b>	Sentences	275.26	57.46	259.08	52.48
	Words	248.86	53.48	249.11	55.70
	Syllable	252.88	54.11	251.68	56.25
<b>Intensity (dB)</b>	Sentences	66.64	1.63	66.66	1.60
	Words	66.72	3.97	66.80	4.01
	Syllable	66.98	4.63	66.95	4.61

## 8. Acknowledgements

In ending not only this thesis, but also my time as a Master's student, I would like to thank some people for their help and support.

First, thank you to all members of the Language Development department for teaching me so much about child language research. In particular, I would like to thank Jeroen Geerts for his support in creating the stimuli and Caroline Rowland for being an incredible PI. I would also like to express my gratitude to Daphne, Inge, Jefta and Sam for their help collecting the data. Most importantly, my biggest thank you goes to my supervisors Melis and Tineke. Whenever I needed help or felt stuck with anything, I knew I could count on both of you for your advice. I truly loved working on this project and I feel very lucky to have had such kind and reliable supervisors from whom I could learn so much.

Thank you also for all the friends I made here in Nijmegen (you all know who you are). I am very grateful for all the fun memories we share, but also for the fact that you never got tired of listening to me practice my talks or helping me with this thesis. A special shout-out also goes to Ryan Law for helping me manage science, oscillations, PhD search and just life in general. Knowing that your office was 20 meters from mine made my internship a lot more fun.

Besides all the amazing people in Nijmegen, I wish to thank my family for supporting me in every way possible over these last years. In particular, I would like to dedicate this thesis to my grandmother, Anna Christina (known to her family as vovó Babi). I have no doubt in my mind that she is the reason for my passion for science.

Last, but certainly not least, I would like to thank all the parents and babies that agreed to participate in this study. Without you, none of this would have been possible.