Received: 12 February 2024 | Revised: 4 July 2024 | Accepted: 20 July 2024

DOI: 10.1111/ejn.16492

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

The effect of visual speech cues on neural tracking of speech in 10-month-old infants

Melis Çetinçelik[1,2,3] | **Antonia Jordan-Barros**[4,5] | **Caroline F. Rowland**[1,6] | **Tineke M. Snijders**[1,3,6]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[2]Department of Experimental Psychology, Utrecht University, Utrecht, The Netherlands

[3]Cognitive Neuropsychology Department, Tilburg University, Tilburg, The Netherlands

[4]Centre for Brain and Cognitive Development, Department of Psychological Science, Birkbeck, University of London, London, UK

[5]Experimental Psychology, University College London, London, UK

[6]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Correspondence
Melis Çetinçelik, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.
Email: melis.cetincelik@mpi.nl

Funding information
Max Planck Society

Edited by: Edmund Lalor

Abstract

While infants' sensitivity to visual speech cues and the benefit of these cues have been well-established by behavioural studies, there is little evidence on the effect of visual speech cues on infants' neural processing of continuous auditory speech. In this study, we investigated whether visual speech cues, such as the movements of the lips, jaw, and larynx, facilitate infants' neural speech tracking. Ten-month-old Dutch-learning infants watched videos of a speaker reciting passages in infant-directed speech while electroencephalography (EEG) was recorded. In the videos, either the full face of the speaker was displayed or the speaker's mouth and jaw were masked with a block, obstructing the visual speech cues. To assess neural tracking, speech-brain coherence (SBC) was calculated, focusing particularly on the stress and syllabic rates (1–1.75 and 2.5–3.5 Hz respectively in our stimuli). First, overall, SBC was compared to surrogate data, and then, differences in SBC in the two conditions were tested at the frequencies of interest. Our results indicated that infants show significant tracking at both stress and syllabic rates. However, no differences were identified between the two conditions, meaning that infants' neural tracking was not modulated further by the presence of visual speech cues. Furthermore, we demonstrated that infants' neural tracking of low-frequency information is related to their subsequent vocabulary development at 18 months. Overall, this study provides evidence that infants' neural tracking of speech is not necessarily impaired when visual speech cues are not fully visible and that neural tracking may be a potential mechanism in successful language acquisition.

KEYWORDS
audiovisual speech, neural tracking of speech, speech-brain coherence, visual speech cues, vocabulary development

Abbreviations: AV, audiovisual; COVID-19, coronavirus disease 2019; EEG, electroencephalography; ERP, event-related potential; ICA, independent component analysis; SBC, speech-brain coherence; TBW, temporal binding window; (m)TRF, (multivariate) temporal response function; (N-)CDI, (Dutch version of the) MacArthur-Bates Communicative Development Inventories.

1 | INTRODUCTION

Speech perception is a multimodal process that incorporates multisensory information from auditory, visual, and/or tactile cues. However, speech perception is commonly studied by taking only the auditory modality into account. This poses a problem for our understanding of how infants learn to process speech, as infants rarely hear the auditory signal in isolation, but are usually presented with speech that is accompanied by other cues in face-to-face interactions with their caregivers. Important cues in such interactions are the visual speech cues that are produced by the rhythmic movements of the speaker's mouth, lips, and larynx, which occur simultaneously with the auditory signal. This means that redundant, but complementary, information is available in the auditory and visual modalities.

Adult listeners benefit from seeing visual speech cues when processing speech, especially in noisy or challenging listening conditions, such as the presence of competing talkers (Rudmann et al., 2003; Zion Golumbic, Cogan, et al., 2013; Zion Golumbic, Ding, et al., 2013) or speech-in-noise (Bernstein et al., 2004; Crosse et al., 2016; Grant & Seitz, 2000; Moradi et al., 2013; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollack, 1954). For instance, viewing a talker's face is linked to decreased effort while listening to speech-in-noise (Fraser et al., 2010), which might be particularly beneficial for hard-of-hearing individuals (Mishra et al., 2014). Note, however, that there are substantial individual differences in the degree with which listeners benefit from seeing additional visual speech information (Bernstein et al., 2000; Tye-Murray et al., 2014).

Infants also show an early sensitivity to the presence of audiovisual (AV) cues and their correspondence with the auditory signal (Kuhl & Meltzoff, 1982; Kushnerenko et al., 2008; Patterson & Werker, 1999, 2003). Even newborns can detect which point-line display of talking faces matches the utterance they hear (Guellaï et al., 2016). Already at 2 months, infants preferentially look at the face articulating the vowel they are hearing over a face producing a mismatched vowel (Patterson & Werker, 2003), an effect which has also been shown for 4.5-month-old infants (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Furthermore, the McGurk effect (McGurk & Macdonald, 1976), a multisensory illusion resulting from AV integration, has been observed at four months of age (Burnham & Dodd, 2004). Neuroimaging studies further suggest that AV sensitivity emerges early in infancy, showing differential event-related potentials (ERPs) to congruent and incongruent utterances presented as matching and non-matching auditory and visual input (Hyde et al., 2011; Kushnerenko et al., 2008, 2013; Reynolds et al., 2014).

The AV benefit also applies to speech perception in infants and children. For instance, Hollich et al. (2005) found that 7.5-month-old infants could segment words presented in continuous speech-in-noise when the stimuli were presented by a talking face, and when they saw an oscilloscope synchronised with speech, but not in an auditory-only (AO) condition where speech was simultaneously presented only with a static picture of the speaker's face. Teinonen et al. (2008) demonstrated that congruent visual articulations of phonemes contributed to 6-month-old infants' learning of phoneme boundaries, whereas no learning was observed with incongruent cues. Similarly, children's phoneme and word recognition abilities were enhanced in AV presentations compared to audio-only presentations in both speech-in-noise and without noise (Lalonde & Holt, 2015, 2016; Lalonde & McCreery, 2020; Maidment et al., 2015; Ross et al., 2011).

Even though infants are sensitive to AV correspondence in speech from an early age, multisensory abilities continue to develop in childhood. Lalonde and Werner (2019) showed that 6- to 8-month-old infants' detection of target syllables in speech-in-noise was better with AV speech compared to AO speech, similar to adults'. However, for the more complex task of speech discrimination, infants relied on simpler visual onset-offset cues, whereas adults used more fine-grained visual cues, indicating that infants are still maturing in their ability to use a wider range of spectrotemporal cues in AV speech perception (Lalonde & Werner, 2019). While children aged 6 to 12 years show similar benefits to adults in simpler tasks like detecting single syllables in noisy speech, this advantage diminished in the more complex speech recognition task (Lalonde & McCreery, 2020). Thus, the development of multisensory abilities in infants extends into childhood and adolescence (Ross et al., 2011; Wallace et al., 2020) which means that, particularly for processing of more complex linguistic stimuli, an adult-like AV benefit cannot be readily assumed.

What are the mechanisms that underlie an AV benefit in speech perception? One potential mechanism is related to the temporal correspondence between auditory and visual information. In AV speech, multiple cross-modal cues subserve this temporal correspondence, and can facilitate speech processing by enhancing temporal expectancy and prediction of auditory speech (Grant & Seitz, 2000; Lalonde & Werner, 2021). For example, an important potential cue comes from the cross-modal correlations between the auditory speech envelope and the visual speech cues provided by the speaker's articulatory movements. There is a temporal correspondence between the opening and closing of the lips and the speech envelope, particularly at the syllable rate (2–7 Hz in adult-directed

speech; Chandrasekaran et al., 2009). While some evidence suggests that articulatory movements tend to precede auditory speech by 100–300 ms (Chandrasekaran et al., 2009), Schwartz and Savariaux (2014) found that the correspondence in continuous speech is actually more varied, ranging between 40 ms audio lead and 200 ms audio lag. However, they also noted that the predictability of AV speech does not necessarily depend on visual input leading the auditory counterpart (Schwartz & Savariaux, 2014). Thus, seeing congruent visual speech cues in addition to hearing auditory speech may allow listeners to generate reliable temporal predictions about the auditory speech signal, and thereby result in more efficient speech processing.

A mechanism that has been suggested to play a key role in speech processing and may be implicated in the AV speech advantage is neural tracking of speech. Neural tracking refers to the synchronisation or phase-locking of neural oscillations to the acoustic dynamics of speech, and involves processes that facilitate effective encoding and processing of auditory input (Doelling et al., 2014; Luo & Poeppel, 2007; Peelle & Davis, 2012). The salient, inherently rhythmic patterns in speech, such as the temporal regularities of syllable and stress information, allows the neural oscillations of the brain to align with the phase of the external input through phase-locking (Giraud & Poeppel, 2012). Key to the potential benefit of audio-visual speech is the fact that visual speech input is also tracked in the brain, with listeners' low-frequency oscillations in visual and auditory cortices, especially in the theta frequency (4–8 Hz, corresponding to syllable rate), being modulated by the speaker's lip movements (Biau et al., 2021). Importantly, oscillations in the visual cortex interact with oscillations in the auditory cortex through connections between the visual and auditory sensory cortices and subcortical pathways (Luo et al., 2010). Given that visual signals often precede their auditory counterpart (Chandrasekaran et al., 2009), this cross-modal integration between sensory cortices can enable auditory oscillations to align their high excitability phase with the expected upcoming input more efficiently compared to the absence of visual information (Luo et al., 2010; Zoefel, 2021). This allows the listener to form temporal predictions about the upcoming speech input, especially the timing of auditory information, thereby optimising speech processing through reliable neural tracking (Peelle & Davis, 2012; Peelle & Sommers, 2015; Schroeder et al., 2008; Zoefel, 2021). Indeed, studies have suggested that visual speech cues influence neural tracking of continuous speech (Bauer et al., 2020; Biau et al., 2021; Bourguignon et al., 2020; Crosse et al., 2015, 2016; Luo et al., 2010; Peelle & Sommers, 2015; Thézé et al., 2020; Zoefel, 2021). AV speech seems to be particularly useful for speech processing in challenging conditions, such as the presence of background noise or competing speakers (Haider et al., 2023; Zion Golumbic, Cogan, et al., 2013; Zion Golumbic, Ding, et al., 2013).

Successful neural tracking of the speech envelope has also been repeatedly demonstrated in infants and young children (Attaheri, Ní Choisdealbha, Di Liberto, et al., 2022; Cantiani et al., 2022; Çetinçelik et al., 2023; Jessen et al., 2019; Kalashnikova et al., 2018; Menn, Michel, et al., 2022; Menn, Ward, et al., 2022; Ortiz Barajas et al., 2021; Power et al., 2012; Tan et al., 2022), and infants' neural speech tracking at the stress and syllable rate has been linked to their later vocabulary outcomes (Attaheri, Ní Choisdealbha, Di Liberto, et al., 2022; Çetinçelik et al., 2023; Menn, Ward, et al., 2022; Ní Choisdealbha et al., 2023). Yet, only a few studies have investigated whether infants' and children's neural tracking is enhanced in the presence of AV speech cues. We might expect to see a substantial AV benefit in infants who are beginning to acquire their native language, especially since, for them, speech perception might be a challenge even under ideal listening conditions. However, the evidence from the small number of studies so far is mixed. Comparing AO, visual-only (VO), and AV recordings of a speaker repeating a single syllable, Power et al. (2012) found that 13-year-olds' theta-band (~4 Hz), but not delta-band (~2 Hz), entrainment to speech was larger for AV speech compared to VO speech, but not when compared to AO speech. They also reported that the preferred phase of the auditory entrainment at the theta band was modulated by the presence of visual speech cues, suggesting cross-modal phase modulation. Comparing neural tracking of continuous infant-directed speech in AO, VO, and AV presentations in 5-month-olds, 4-year-olds, and adults, Tan et al. (2022) found an auditory–visual speech benefit for the infant and adult groups, meaning that AV speech tracking was more accurate compared to the audio-only and VO modalities combined. However, this benefit was not found for the 4-year-olds, and furthermore, was not correlated with relative attention to the speaker's mouth in the AV condition (Tan et al., 2022). Moreover, while Tan et al. (2022) reported that infants did not track VO continuous speech, Ní Choisdealbha et al. (2024) reported that both 5- and 8-month-olds showed neural tracking of repeated syllables presented VO by a silent talking face, which might be due to differences in stimulus complexity (continuous speech vs. repeated syllables).

Taken together, while some research has been carried out on the role of visual speech cues on infants' and children's neural tracking of speech, the evidence is inconclusive. Furthermore, none of the previous studies tested the effects of visual cues per se, by obstructing visual access to these cues or testing their effects separately, but

rather compared neural tracking in different modalities (AV vs A-only and V-only), which makes it difficult to draw firm conclusions about the role of specific visual speech cues in neural tracking. In other words, it is not possible to conclude that any observed AV-benefit in neural tracking of speech is due to lip movements, because it might also be due to other cues that are not present in AO speech (e.g., head and eyebrow movements) or general attentional enhancement from multimodal redundant stimuli (i.e., cues occurring together in different modalities) (Watson et al., 2014).

In this study, we investigated whether visual speech cues facilitate infants' neural tracking of speech by masking the visual cues that infants might benefit from when processing AV speech. To this end, we tested 10-month-old infants, since there is a developmental shift in infants' looking patterns when viewing a talking face roughly around this age; infants start looking longer at the mouth of a speaker compared to the eyes between 4 and 8 months, up until at least 12 months (Morin-Lessard et al., 2019; Lewkowicz & Hansen-Tift, 2012; Pons et al., 2015; though for counter-evidence see Sekiyama et al., 2021; also see Bastianello et al., 2022 for a systematic review). We presented infants with videos in which the speaker's face was either fully present, or in which the lower half of the face was covered with a static block, occluding the mouth, lips, jaw, and larynx of the speaker. This presentation, to some extent, resembles young infants' interactions with speakers outside of their home environment in the COVID-19 pandemic, when face masks became a part of everyday life. We hypothesised that, in the presence of visual speech cues, infants will show larger neural tracking effects at the stress and syllable frequencies, indexed by higher levels of speech-brain coherence. Speech-brain coherence is a measure of the consistency of the phase difference between the brain oscillations and the speech amplitude envelope at a given frequency, which directly addresses the synchronicity between the brain activity and the speech envelope (Peelle et al., 2013). We predict that these effects might be particularly evident at the stress and syllable frequencies, because at these frequencies, most modulations are seen in the amplitude envelope of infant-directed speech (Leong et al., 2017), and because the articulatory movements are mostly correlated with the speech amplitude envelope at the syllable rate in adult-directed speech, making it the frequency rate at which the largest AV benefit was observed in neural tracking of speech (Chandrasekaran et al., 2009; Crosse et al., 2016). Furthermore, we assessed infants' looking times to investigate whether infants show differences in attention in response to the fully AV condition compared to the blocked videos, which would result in longer looking times to the AV condition. Finally, we assessed whether individual differences in neural speech tracking abilities at stress and syllable frequency rates are related to later vocabulary development by relating neural tracking at 10 months to vocabulary growth between 10 and 18 months. In stress-timed languages such as Dutch, stressed syllables play an important role in word segmentation (Jusczyk et al., 1999), and therefore sensitivity to those cues in continuous speech might be beneficial for infants at this age.

2 | METHODS

The main analysis regarding the effect of visual cues on neural tracking of speech was preregistered (see https://aspredicted.org/blind.php?x=PJL_8H9).[1] The preregistration also includes an analysis of word segmentation, but this is not included in the analysis reported in the current paper. Note that the research question regarding vocabulary development was added after the initial preregistration, but before data analysis started. Furthermore, an exploratory (not preregistered) analysis of infants' electroencephalography (EEG) power at the theta band can be found in Supplementary Materials.

2.1 | Participants

The final participant sample consisted of 34 Dutch-learning infants (21 female, mean age 308.5 days, range: 291–320 days). Twenty-nine other infants were tested but excluded from the current analysis due to not having enough artefact-free trials ($n = 21$), having more than four noisy or flat channels ($n = 3$), technical errors ($n = 3$), and not starting the experiment or not providing any usable data because of fussiness ($n = 2$). Of those 34 infants, 27 also provided the follow-up measures for vocabulary size (CDI) at 18 months. Therefore, the sample for the analysis relating neural tracking to later vocabulary development consisted of 27 infants. Infants

[1]Note that in the preregistration, we specified that frequencies and electrodes with significant speech-brain coherence would be determined using cluster-based permutation tests. However, as this test identified one large cluster over the 1–10-Hz frequency range, we further assessed the stressed syllable and syllable frequencies (determined from acoustic analyses of the stimuli). This assessment is driven by the theoretical reasoning that neural tracking of low-frequency rates is crucial for infants, as low-frequency rates are particularly prominent in infant-directed speech and play an essential role in vocabulary development (Goswami, 2019; Leong & Goswami, 2015), possibly because they provide infants with an advantage in word segmentation. This hypothesis was included in the word segmentation section of the preregistration, which was not included in the current paper.

were born full-term, were typically developing, and came from monolingual Dutch-speaking backgrounds without any neurological or language problems in the immediate family. Among the caregivers of the included participants, 97% had a university or applied university degree (based on highest level of education among the two caregivers). Participants were recruited from the Nijmegen Baby and Child Research Center database. The study was approved by the Ethical Board of Social Sciences, Radboud University, Nijmegen. Caregiver(s) provided written informed consent in accordance with the Declaration of Helsinki. Caregiver(s) were offered a choice between 20 Euros and a book for their participation.

2.2 | Materials

Materials consisted of a continuous AV speech phase (familiarisation phase) and an audio-only test phase in which isolated words were repeated (test phase). In this paper, only the data from the continuous speech phase is analysed. The test phase is not included in the current paper, but the details of the test phase are reported in the Supplementary Materials for completeness. Table 1 provides an example of the continuous speech phase of an experimental block (see Supplementary Materials, Table S2 for the full set of materials, including the repeated words).

2.2.1 | Stimuli

For the continuous speech stimuli, 30 blocks were created, with each block comprising four sentences containing 8–12 syllables. Two versions (A and B) of the continuous speech stimuli were created by using the same sentences with a different repeated word (see Supplementary Materials, Table S2). Note that in order to test the effect of familiarisation in a later test phase (analysis not reported here), in each block, one word was repeated in the continuous sentences, which was a low-frequency trochaic word selected from the Dutch CELEX database (Baayen et al., 1995). The repeated words occurred in the sentence-final position only in the second or third sentence in one block.

The continuous speech stimuli consisted of videos of a female Dutch native speaker reciting the sentences. The actor faced the camera during stimulus recording, looked at a picture of an infant and was instructed to use infant-directed speech. The videos were used as is for the AV condition. For the AV-Blocked (AVb) condition, the videos were further edited so that a static grey rectangle occluded the speaker's lower face including the mouth, jaw, and larynx, so that all articulatory movements were obscured. The speaker's eyes and upper face remained visible at all times, and the acoustic properties of speech were identical across the AV and AVb conditions. The static rectangle always extended to the bottommost edge of the video, creating the appearance that the speaker could be speaking from behind a grey wall or occlusion. The videos were processed using DaVinci Resolve software (version 16; Blackmagic Design, 2020). Figure 1 illustrates a visualisation of the videos in both conditions.

The audio of the continuous speech stimuli was normalised to 70 dB using Praat (Boersma & Weenink, 2021). The continuous speech stimuli had a mean sentence duration of 3201 ms ($SD = 511$ ms) and an inter-sentence interval of approximately 1500 ms. The continuous speech blocks lasted between 16.2 and 20.9 s ($M = 18.79$ s, $SD = 0.98$ s). A comparison of the acoustic properties of the stimuli in versions A and B indicated no significant differences between the two versions (see Supplementary Materials, Table S3). Stimuli were annotated to identify the frequencies of the linguistic units (stressed syllables and syllables) in the stimuli. The duration of all stressed syllables, syllables, words, and sentences were transcribed, and mean frequencies and frequency ranges

TABLE 1 An example of the continuous speech phase of an experimental block (repeated words are in bold, English translations in parentheses).

Continuous speech phase
1. Ik doe altijd wat **dille** op mijn vis. *(I always put some dill on my fish.)*
2. Duitse koks koken graag met **dille**. *(German chefs like to cook with dill.)*
3. Wij hebben **dille** in de tuin. *(We have dill in the garden.)*
4. Zullen we in de salade wat **dille** doen? *(Shall we put some dill in the salad?)*

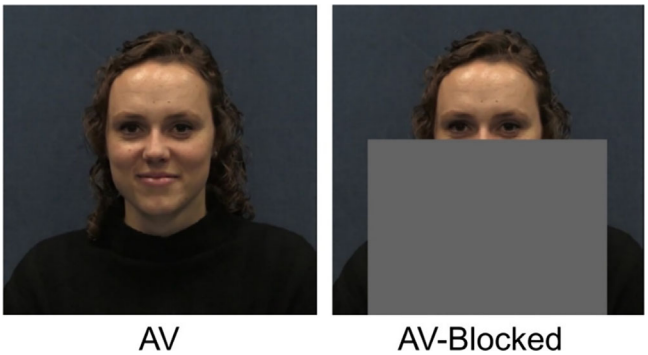

FIGURE 1 Images taken from the video stimuli in both conditions. The audiovisual (AV) condition is illustrated on the left, and the AV-blocked condition is shown on the right.

were calculated for each unit, excluding pauses between sentences but including inter-sentence pauses. Stressed syllables occurred at a rate of 0.9–1.7 Hz ($M = 1.3$ Hz) and syllables at 2.6–3.6 Hz ($M = 3.1$ Hz). The frequency ranges selected for subsequent analyses were 1–1.75 Hz for the stress rate and 2.5–3.5 Hz for the syllable rate, due to the 0.25-Hz frequency resolution of the Fourier transform.

2.2.2 | Language outcome tests

Infants' receptive and expressive vocabulary knowledge was assessed using the short Dutch version of the MacArthur-Bates Communicative Development Inventories (N-CDI; Zink & Lejaegere, 2003). The CDI is a caregiver-report vocabulary checklist in which caregivers indicate the items that their children "understand" and "understand and say," indexing receptive and expressive vocabulary sizes. At the time of the EEG measurement at 10 months, caregivers were asked to fill in the online N-CDI-1 (maximum possible score = 103 for each category). At 18 months, caregivers were invited to fill in the N-CDI-2 to test infants' vocabulary development (maximum possible score = 112 for each category).

2.3 | Design

Participants were presented with 30 experimental blocks consisting of a continuous speech phase (video) and isolated words phase (audio-only, analysis not reported here). In half of the continuous speech videos, the speaker's face was fully visible (AV-condition), whereas in the other half, the speaker's mouth, jaw, and larynx were occluded with a static grey rectangle to cover the visual speech cues (AVb condition). Each participant was presented with videos in both the AV and AVb conditions. The visual speech cue condition (AV vs. AVb) was changed every two blocks, except for the last two blocks in which the condition was switched after one block to ensure that an equal number of AV and AVb videos were presented.

Within each block, each continuous speech video (familiarisation phase, analysed here) was followed by the audio-only presentation of the two isolated words (test phase, not analysed here). The interstimulus interval between the familiarisation and the test phases was approximately 1500 ms, and the inter-trial interval between two experimental blocks was approximately 3000 ms. AV attention getters with moving images and simple repeating sounds were played between every two blocks in a pseudo-randomised order to maintain infants' attention on the screen. The order of presentation of blocks was pseudo-randomised and counterbalanced across participants using eight different experimental lists, with half of the lists using the words from version A as the target words, and the other half using the version B target words.

2.4 | Procedure

Each session was run by two experimenters. One experimenter explained the study procedure to the caregiver while the infant played on the playmat, and the other experimenter pre-gelled the EEG cap to minimise setup time, then fitted the cap on the infant's head and added more gel if necessary. Next, the infant was seated in their caregiver's lap in a sound-attenuated and electrically shielded experiment booth, approximately 70 cm away from a 24-in. monitor. Videos were presented at the centre of the screen (20 cm × 20 cm), and audio was played in stereo over two loudspeakers at approximately 65 dB. Experimental stimuli were presented using Presentation (Version 20.2, Neurobehavioral Systems, Inc).

The experiment began with an attention-getter, followed by two 10-s videos of the speaker gazing silently at the infant, one video from the AV-condition and one from the AVb condition, to get the infant accustomed to the face of the speaker. Afterwards, the presentation of the experimental blocks was initiated. Caregivers were asked to listen to masking music through noise-cancelling headphones and not to interact with the infant during the session. Infants were offered silent toys or breadsticks if they became restless. The experimenters ran the experiment and the EEG acquisition from outside of the experiment booth, and observed the participants via a video link. The sessions were video recorded for offline coding of the infants' looking behaviour. If the infant became fussy during the experiment, a short break was taken, and a silent cartoon video was played. The experiment was concluded if the infant became distressed or stopped attending to the screen. The experiment lasted approximately 15 min, and the whole session (including preparation and capping) lasted from 45 min to an hour. The COVID-19 regulations of the institution were followed during testing, and caregivers and experimenters wore facemasks for the first eight participants tested, of which seven are included in the speech-brain coherence analysis, and two are included in the analysis relating SBC to later language outcomes. Removing these participants from the analyses generated the same pattern of results with the remainder; therefore, they were included in the final sample.

2.4.1 | EEG recordings

EEG data were recorded with a 32-channel EEG system (actiCAP with active Ag/AgCl electrodes), BrainAmp DC amplifier, and Brain Vision Recorder software (Brain Products GmbH, Germany). The electrodes were placed in accordance with the extended 10–20 system, and EEG was recorded from Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, and P8. Two additional electrodes were placed directly on the mastoid bones ("TP9L", "TP10L") in addition to the mastoid electrodes in the cap (TP9, TP10) as potential reference electrodes. Vertical eye movements were captured using the electrode above (Fp1) and an additional electrode placed below the left eye, and horizontal eye movements using the two electrodes at the outer canthi of the eyes (FT9, FT10). FCz served as the online reference. Data were recorded at a sampling rate of 500 Hz, with an online time cut-off of 10 s and a high cut-off of 1000 Hz. Electrode impedances were typically kept under 25 kΩ.

2.5 | Data processing

2.5.1 | Gaze measures

Infants' looking behaviour were manually coded using ELAN (version 6.3; 2022) for the duration of the continuous speech videos. To this end, infants' looks to the screen and looks away from the screen were coded frame-by-frame. Afterwards, the gaze data were segmented into 4-s epochs with a 1-s sliding window, to match them with the EEG epochs. Infants' attention in each trial was computed by obtaining the proportion of looking time to the screen per four-second epoch. Epochs were marked for exclusion from the speech-brain coherence analysis if the infant's attention towards the video was less than 25% in a given 4-s epoch (similar to Tan et al., 2022). We also conducted a control analysis with a higher threshold of 50% looking. This yielded the same pattern of results, though with a smaller number of included participants. That analysis is reported in the Supplementary Materials.

2.5.2 | EEG measures

EEG pre-processing and analysis was done using the Fieldtrip toolbox for EEG/MEG-analysis (Oostenveld et al., 2011) in MATLAB (version 2020b). As a first pre-processing step, the complete dataset was pre-processed to provide as much data as possible for the independent component analysis (ICA; Makeig et al., 1996). Data were filtered with a Butterworth high-pass filter at 0.1 Hz (−12 dB/oct) and low-pass filter at 30 Hz. Continuous data were segmented into 1-s epochs for artefact rejection an were visually inspected to remove bad channels (channels with excessive noise, flat lines, or electrode drifts) and epochs containing high-amplitude artefacts (150 μV for EEG channels, 250 μV for EOG channels). After the removal of large artefacts, data were decomposed using Infomax ICA (Bell & Sejnowski, 1995), and eye movement and single-channel noise components were identified by visual inspection of the data and the component morphology which were then removed in the next step.

Afterwards, raw EEG data were re-segmented into 4-s epochs using a 1-s sliding window for the speech-brain coherence analysis, with each first trial time-locked to the onset of the continuous speech video and continuing until the end of the video, ending with the last full 4-s epoch. Raw data were again filtered at 0.1 and 30 Hz, and ocular and single-channel noise components identified with ICA were removed from the epoched data (mean number of rejected eye components = 3.3, mean number of single-channel noise components = 4.5). Two posterior channels (P7 and P8) were removed because they were too noisy across many datasets. Data were then re-referenced to the linked mastoids (TP9L and TP10L, or TP9 and TP10, or a bilateral combination), and epochs were demeaned. If the linked mastoid was identified as noisy, re-referencing was performed to a single mastoid (for one infant).

The acoustic envelope of the speech stimuli of the continuous speech stimuli was computed using a Hilbert-transform with a second-order Butterworth filter and downsampled to 500 Hz to match the sampling rate of the EEG signal. The speech envelope was then cut into 4-s snippets with a 1-s sliding window, and was combined with the clean, epoched EEG data. This resulted in 13 to 17 trials per video, depending on the video length, and a maximum number of 455 or 460 epochs overall, with 225–232 trials in each condition, depending on the experimental list. Remaining artefacts exceeding ±150 μV in the 4-s epochs, as well as trials which did not pass the looking-time criterion (<25% attention during each 4-second trial), were excluded.

Infants with at least 30 artefact-free epochs in each condition (60 overall) were included in the final dataset, to ensure a reliable coherence estimate (Bastos & Schoffelen, 2016). On average, infants had a total of 135.9 included trials ($SD = 54$; $M_{AV} = 72.5$, $SD_{AV} = 32.9$; $M_{AV\text{-}Blocked} = 63.4$, $SD_{AV\text{-}Blocked} = 25$). Missing EEG channels were repaired (mean number of repaired channels = 0.7, range = 0–3) using spherical spline interpolation (Perrin et al., 1989).

The EEG data and speech envelope were Fourier-transformed between 1 and 10 Hz with a frequency resolution of 0.25 Hz, given the 4-s trials. The coherence between the speech envelope and the EEG signal was computed for each channel-speech signal combination, resulting in a coherence value between 0 and 1 which reflects the consistency of the phase difference between the two signals, amplitude envelope and brain activity, at a given frequency (Peelle et al., 2013). Coherence was calculated using the following formula, where $S_{xy}(\omega)$ is the cross-spectral density between the speech envelope (x) and the EEG signal (y) at frequency ω, and $S_{xx}(\omega)$ and $S_{yy}(\omega)$ are the power spectra of the speech envelope and the EEG signal (Bastos & Schoffelen, 2016; Rosenberg et al., 1989):

$$\mathrm{coh}_{xy}(\omega) = \frac{|S_{xy}(\omega)|}{\sqrt{S_{xx}(\omega)S_{yy}(\omega)}}$$

2.6 | Analysis

2.6.1 | Speech-brain coherence

To establish the presence of neural speech tracking (i.e., whether speech-brain coherence values were higher than above-chance level), observed coherence was compared against surrogate data. This surrogate data was created by shuffling the speech amplitude envelope across epochs, that is, randomly pairing the EEG trials with the envelope epochs to remove any temporal correspondence between the EEG and the envelope in the surrogate data, and computing the average coherence of the shuffled envelope and the EEG data for 100 permutations, regardless of the experimental condition (Vanden Bosch der Nederlanden et al., 2022).

Observed and surrogate coherence values, both overall and in each experimental condition, were compared using non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007) over all EEG electrodes (cluster alpha level = .05, number of permutations = 1000 or 10,000 if the initial test yielded a p-value close to the significance threshold, minimum number of neighbouring channels = 1, all reported tests p-corrected). Analyses were first conducted for the whole frequency range between 1 and 10 Hz, and then separately for the stimulus-driven frequency ranges of interest, which are the stressed syllable (1–1.75 Hz) and syllable frequencies (2.5–3.5 Hz). To test the effect of the visual speech cues, speech-brain coherence in the two experimental conditions (AV vs. AVb) were compared in both the stressed syllable and syllable frequency ranges, over all electrodes. Finally, for subsequent analyses, speech-brain coherence (overall and per condition) was z-score transformed using the mean and the standard deviation of the surrogate data in the respective condition to correct for coherence bias (Bastos & Schoffelen, 2016; see Vanden Bosch der Nederlanden et al., 2022 for a similar approach). Then, the z-transformed coherence values over the included electrodes in the frequency bands of interest (stressed syllable and syllable) were averaged, resulting in one coherence value per frequency band for each participant.

2.6.2 | Looking behaviour and EEG power

Infants' mean looking times over the 4-s epochs were compared with paired-samples t-tests to examine differences in infants' attention to the stimuli. This was conducted for the included 4-s epochs, as well as over all the trials, regardless of inclusion in the EEG analysis. Additionally, differences in attention in response to the two conditions were examined with an exploratory analysis of absolute EEG power (see Supplementary Materials C for details).

2.6.3 | Language tests

Infants' receptive and expressive vocabulary scores at 10 and 18 months were calculated using the items caregivers ticked on the N-CDI. Proportion scores in each category were calculated by dividing the individual score by the total number of items per measure in order to have the scores on a 0–1 scale.

To assess the relationship between neural speech tracking abilities and subsequent vocabulary development, linear regression models were fit to infants' receptive and expressive vocabulary at 18 months, using infants' z-score transformed speech-brain coherence values at the stressed syllable and syllable rates and their receptive or expressive vocabulary scores at 10 months as predictors, to control for vocabulary size at the time of the EEG measurement. These analyses were conducted using the *stats* package in R (version 4.2.2; R Core Team, 2022).

3 | RESULTS

We first present results from a looking time analysis that tested for differences in infants' mean looking times in the included trials in the AV and AVb conditions. Then, we introduce the results of the speech-brain coherence analyses, first testing for the presence of speech-brain coherence, and then comparing speech-brain coherence in the two conditions. Finally, we present the results

linking individual differences in neural tracking at 10 months to future vocabulary development. Note that we also conducted an exploratory analysis of infants' absolute EEG theta power to test for attention differences in response to the two conditions, where attentional differences would be reflected in greater EEG power at the theta band (Begus & Bonawitz, 2020; Orekhova et al., 2006). This analysis did not reveal any significant differences between the two conditions (see Supplementary Materials C).

3.1 | Looking times

Infants' mean looking times in the included 4-s trials were significantly longer in the AV condition ($M = 3.54$ s [88.44%], $SD = 0.31$ s [7.78%]) compared to the AVb condition ($M = 3.22$ s [80.54%], $SD = 0.39$ s [9.66%]), $t(33) = 7.05$, $p < .001$). Figure 2 illustrates the infants' mean percentage looking time in each condition (see Supplementary Materials, Figure S1 for infants' attention over all trials, regardless of inclusion).

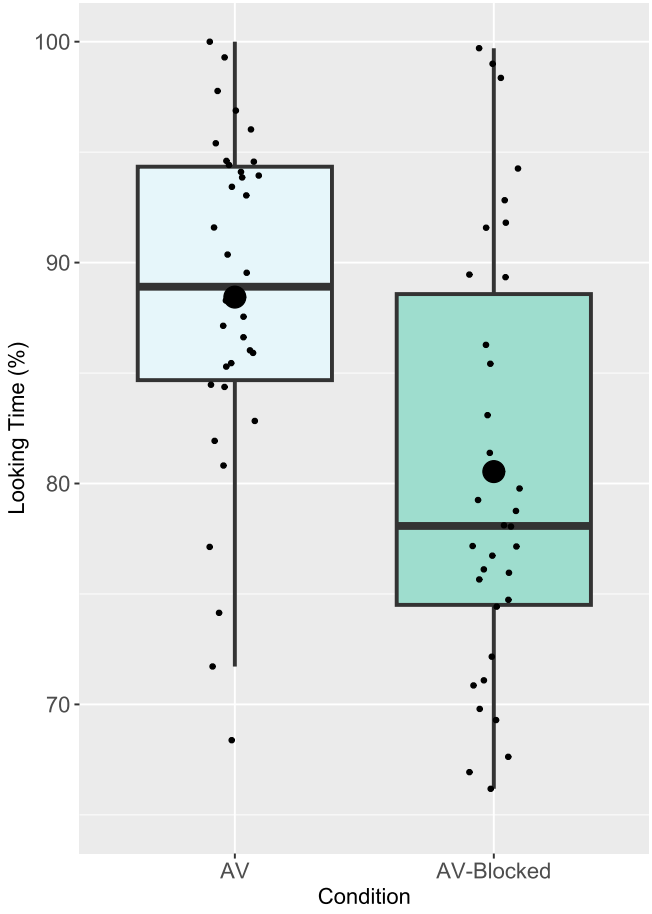

FIGURE 2 Proportion of infants' mean looking times in the included trials (percentages, calculated by the mean proportion of looking to the screen during the 4-s epochs) in the audiovisual (AV) and AV-blocked conditions. The length of the boxes represents the interquartile range (IQR), the whiskers show the lower and upper values within 1.5 IQR, the small points show each infants' mean proportion of looking averaged over trials, and the large black points denote the mean proportion of looking across participants in each condition.

3.2 | Speech-brain coherence

First, speech-brain coherence in the observed data was compared to surrogate data to test for the presence of neural tracking in the 1–10-Hz frequency range, without averaging over electrodes or frequencies, with a cluster-based permutation test. This comparison yielded one large cluster with a significant difference between the observed and surrogate data (cluster $p_{corrected} = .002$) that included all electrodes and all frequencies. The 1–10-Hz frequency range covers the stressed syllable, syllable, and word rates. Next, observed overall coherence was compared to surrogate data in the stimulus-driven frequency ranges of interest, which are the stress (1–1.75 Hz) and syllable (2.5–3.5 Hz) frequencies. In both frequency ranges, speech-brain coherence was significantly higher in the observed data than in the surrogate data, and both tests resulted in clusters that encompassed all electrodes ($p_{corrected} = .002$). Figure 3 illustrates the scalp topography of overall coherence and the coherence values in comparison to surrogate data.

Then, neural tracking was assessed separately for the two conditions (AV and AVb), by comparing coherence in the observed data against the surrogate data in both conditions in the two frequency ranges of interest. Again, one large positive cluster was identified in each range for both conditions ($p_{corrected} = .002$).

Finally, we tested for differences between the AV and AVb conditions to assess whether visual speech cues modulated infants' speech-brain coherence by comparing the observed coherence in the two conditions. First, we assessed the whole frequency range between 1 and 10 Hz, which resulted in 1 positive and 6 negative clusters, but the differences were not significant after correction (p of first positive cluster $= 1$; p of first negative cluster $= .10$; 10,000 permutations. See Supplementary Materials, Figure S3 for an illustration of the probability data for each channel-frequency point). Second, as we predicted that the differences would be most prominent at the stressed syllable and syllable frequencies, we compared coherence values separately at these specific frequency rates. No clusters were identified in either frequency range. Figure 4 shows the coherence distributions and the difference in coherence between the two conditions.

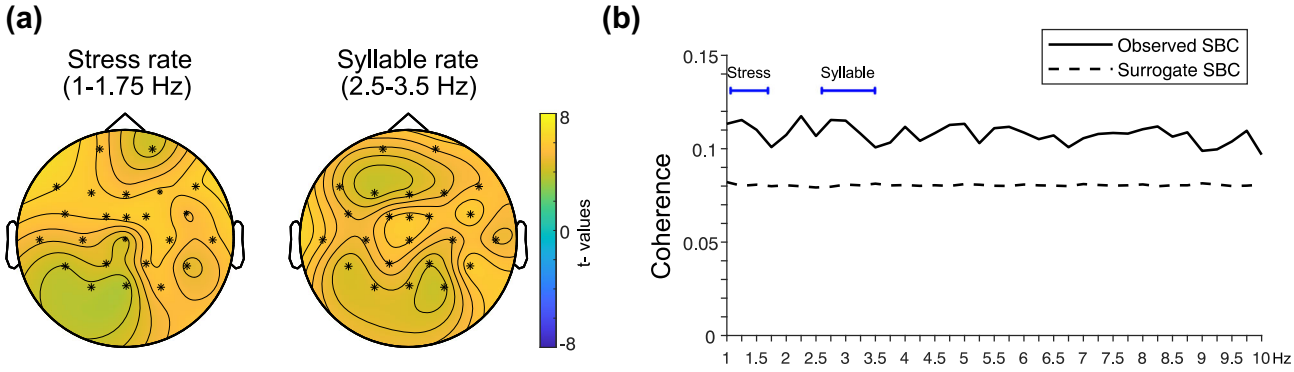

FIGURE 3 (a) Scalp topography of overall speech-brain coherence (SBC) in the stress and syllable rates, showing the *t*-values of the cluster-based permutation test comparing observed SBC to surrogate data. Electrodes that are included in the clusters are highlighted with asterisks. (b) SBC in overall and surrogate data, averaged over all electrodes that form the cluster.

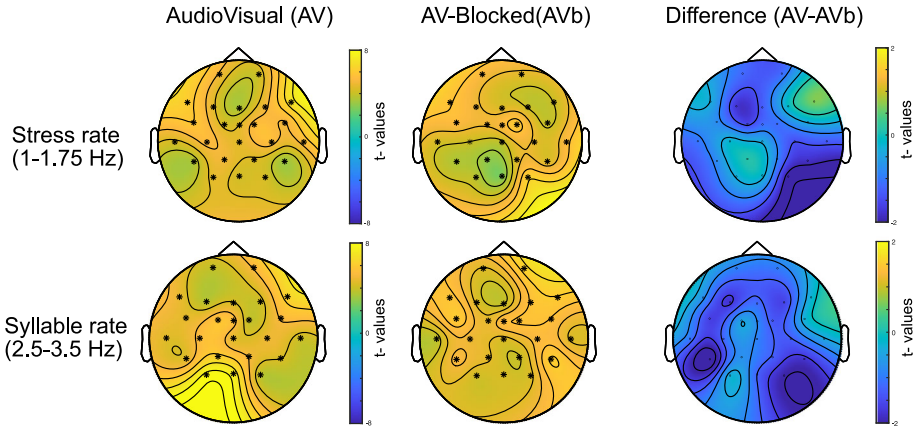

FIGURE 4 Left and middle columns: scalp topography of speech-brain coherence in the stress and syllable rates, in the audiovisual (AV) and AV-blocked (AVb) conditions, showing the *t*-values of the cluster-based permutation test comparing observed SBC to surrogate data. Right column: scalp topography of the speech-brain coherence showing the difference between the AV and AVb conditions. Note that different scales were used for the colour bars for SBC in the AV and AVb conditions (*t*-values between −8 and 8) and the difference figure (−2 and 2) to illustrate differences better. Electrodes that are included in the clusters of the AV and AVb conditions are highlighted with asterisks.

3.3 | Neural tracking and vocabulary development

To assess whether infants' neural speech tracking was related to their later vocabulary development, regression models were fit separately to the receptive and expressive vocabulary outcomes at 18 months, using the *z*-score transformed SBC values at the stressed syllable and syllable rates as predictors. We included infants' 10-month receptive and expressive vocabulary development as predictors in the respective models, to control for concurrent vocabulary at 10 months. The proportion scores for the receptive and expressive CDI measures were entered as the outcome variables in each model. While some predictors showed moderate correlations (Table 2), variance inflation factors (VIFs) suggested that multicollinearity did not occur for either model (VIFs < 1.26).

The results of the receptive vocabulary model indicated a significant model fit, $F(3,23) = 14.39$, $p < .001$ (see Table 3). Infants' neural tracking at the syllable rate at 10 months significantly predicted their receptive vocabulary at 18 months ($\beta = 0.08$, $SE = 0.03$, $t = 2.57$, $p = .017$), after controlling for 10-month receptive vocabulary, which was also a significant predictor of 18-month receptive vocabulary ($\beta = 0.91$, $SE = 0.16$, $t = 5.80$, $p < .001$). Infants' stress-rate tracking was not a significant predictor of later receptive vocabulary.

Then, we fitted a regression model to the 18-month expressive vocabulary data. Testing the model assumptions suggested heteroscedasticity of the model, and therefore, the outcome variable was square-root transformed to fix heteroscedasticity (non-constant variance $p = .008$). Thus, the fitted model included the square-root transformed expressive vocabulary score at 18 months

TABLE 2 Means, standard deviations, and correlations with confidence intervals of the predictor variables in both models. Proportion values are reported for the vocabulary measures.

Variable	*M*	*SD*	1	2	3
1. SBC stress rate	1.20	.93			
2. SBC syllable rate	1.23	.88	.44* [.07, .70]		
3. Receptive voc. 10 months	.20	.15	−.07 [−.43, .32]	−.02 [−.39, .37]	
4. Expressive voc. 10 months	.01	.02	.02 [−.36, .40]	−.11 [−.47, .28]	.44* [.08, .70]

Note: Values in square brackets indicate the 95% confidence interval for each correlation.
*$p < .05$. 1: Speech-brain coherence at the stress rate (*z*-score). 2: Speech-brain coherence at the syllable rate (*z*-score). 3: Receptive vocabulary at 10 months.

TABLE 3 The results of the linear regression models with the *z*-score transformed speech-brain coherence (SBC) at the stress and syllable rates and 10-month receptive vocabulary (proportion) as predictors, and receptive vocabulary at 18 months (proportion) as the outcome measure.

Receptive vocabulary at 18 months				
Predictors	***Estimate***	***SE***	***CI***	***p***
(intercept)	.23	.06	.12–.35	<.001***
SBC stress rate	.02	.03	−.04–.08	.514
SBC syllable rate	.08	.03	.01–.14	.017*
Receptive voc. 10 months	.91	.16	.59–1.24	<.001***
R^2/R^2 adjusted	.652/.607			

***$p < .001$, *$p < .05$.

(proportion) as the outcome variables, and the *z*-score transformed SBC values at the stress and syllable rates, and the 10-month expressive vocabulary score (proportion) as predictors. This model also suggested a significant model fit, $F(3,23) = 9.99$, $p < .001$ (see Table 4). However, this time, infants' stress-rate tracking was a significant predictor of later expressive vocabulary ($\beta = 0.08$, $SE = 0.03$, $t = 2.75$, $p = .011$), controlling for 10-month-old vocabulary. Again, infants' 10-month expressive vocabulary also further predicted their later vocabulary size ($\beta = 4.90$, $SE = 1.09$, $t = 4.48$, $p < .001$).

4 | DISCUSSION

This study investigated the role of visual speech cues on preverbal infants' neural speech processing. Ten-month-old infants' neural tracking of speech, assessed by speech-brain coherence, was analysed to examine whether occluding a speaker's articulatory movements of the mouth, lips, jaw, and larynx results in reduced neural tracking of speech. We especially focussed on the stressed syllable and syllable frequencies, which are key for speech segmentation (Jusczyk et al., 1999). Infants' looking times in each condition were compared to test for differences in attention. Additionally, infants' neural tracking was assessed in relation to their vocabulary development to determine whether infants with better neural tracking abilities at 10 months have a larger vocabulary size at 18 months.

In line with our expectations, infants showed reliable neural tracking of continuous AV speech at 10 months. Neural speech tracking was observed in the overall 1–10-Hz frequency range. Moreover, at both the stress and syllable rate frequencies, we observed significant speech-brain coherence, suggesting that the infant brain takes over the rhythm of speech and tracks the speech input at the frequencies that are most relevant and pronounced in infant-directed speech; the stressed syllable and syllable rates (Leong & Goswami, 2015). Thus, tracking the amplitude modulations at these frequency rates may give infants an advantage in extracting linguistic units from the continuous speech stream (Goswami, 2019). Neural tracking was observed in both speech that was fully AV (AV-condition), in which the speaker's face was fully visible, and in the AVb condition, where the speaker's visual speech cues

TABLE 4 The results of the linear regression models with the *z*-score transformed speech-brain coherence (SBC) at the stress and syllable rates, and 10-month expressive vocabulary (proportion) as predictors, and expressive vocabulary at 18 months (proportion; square-root transformed) as the outcome measure.

Expressive vocabulary at 18 months				
Predictors	***Estimate***	***SE***	***CI***	***p***
(intercept)	.22	.05	.13–.32	<.001***
SBC stress rate	.08	.03	.02–.14	.011*
SBC syllable rate	−.03	.03	−.10–.03	.283
Expressive voc. 10 months	4.90	1.09	2.64–7.16	<.001***
R^2/R^2 adjusted	0.566/0.509			

***$p < .001$, *$p < .05$.

were occluded with a block. These results are consistent with other recent findings demonstrating infants' neural tracking of AO and AV speech (Attaheri, Ní Choisdealbha, Di Liberto, et al., 2022; Cantiani et al., 2022; Çetinçelik et al., 2023; Jessen et al., 2019; Kalashnikova et al., 2018; Menn, Michel, et al., 2022; Menn, Ward, et al., 2022; Ní Choisdealbha et al., 2023; Ortiz Barajas et al., 2021; Power et al., 2012; Tan et al., 2022).

We also predicted that infants' neural tracking of speech would be stronger when visual speech cues are visible in AV speech. Our results, however, did not support this prediction. We observed significant neural tracking of speech at the stress and the syllable rate, which was not modulated by the presence of visual speech cues. This finding is in line with the results for the 4-year-old children in Tan et al. (2022), although it contrasts with their results for the 5-month-olds.

However, we did find that infants attended to the screen more in the fully AV condition compared to the AVb condition, indicating that having access to visual cues from the speaker may influence infants' attention to a talking face. While eye-tracking data has not been recorded in the current study, and thus, it is not possible to ascertain whether infants were looking preferentially at the speaker's mouth or the eyes, the looking time results suggest that seeing a speaker's visual speech cues in addition to hearing auditory speech input leads infants to attend longer to a speaker, and therefore possibly receive more input from the speaker, with greater attention. Yet, the attention advantage in the AV condition did not translate into a neural tracking advantage, also consistent with Tan et al. (2022). While the 4-year-olds in Tan et al. (2022) also attended to the screen significantly more in the AV condition compared to the AO and VO conditions, they did not show an AV-benefit in neural tracking. In contrast, the 5-month-olds in Tan et al. (2022) looked equally long in the AV and AO conditions (but attended more in the AV compared to the VO condition), but nevertheless had greater prediction accuracy in the AV condition. The 5-month-old infants' preferential attention to the speaker's mouth was only correlated with the accuracy of VO speech tracking, but not with AV speech tracking (Tan et al., 2022).

Our results indicating comparable levels of neural speech tracking in the AV and AVb conditions were surprising. There are several possible explanations for this result. One is that 10-month-old infants cannot yet completely make use of redundant visual information for neural tracking of continuous speech. Using visual information to facilitate speech processing may require both general perceptual mechanisms and speech-specific mechanisms stemming from phonetic and lexical knowledge, which are still developing in childhood (Lalonde & Werner, 2021). The integration of complex temporal correspondences in continuous speech, such as the cross-modal correspondences between the visual and auditory speech envelope may require more advanced speech-specific mechanisms, which 10-month-old infants might not have access to yet. The development of the AV-benefit in neural tracking might be further related to the development of the temporal binding window (TBW). The TBW represents an AV temporal integration ability, with a narrower TBW reflecting a higher accuracy. Infants' TBW is relatively large compared to adults', starting to narrow only around 4 years of age and continues gradually narrowing over the course of childhood (Lewkowicz, 2010; Lewkowicz & Flom, 2014). As neural tracking of AV speech requires online perception and integration of multisensory information, it is possible that it necessitates a narrow-enough TBW so that listeners can fully benefit from cross-modal information. Though note that this explanation is contrary to the findings of Tan et al. (2022) who observed an AV speech benefit already with 5-month-old infants.

An alternative explanation is that 10-month-old infants are indeed sensitive to the visual speech cues in

AV speech, but they do not require them for successful speech processing, at least in noise-free conditions. More specifically, in the case of neural tracking, infants might not depend on additional cues to identify the stress and syllable-rate phase onsets in speech. Indeed, previous research has suggested that infants' and newborns' phase tracking was unaffected by their language experience, suggesting that phase tracking may be a universal mechanism that can be observed even when infants are not overtly attending to speech (Ortiz Barajas et al., 2021). Taken together with the results of Tan et al. (2022), our results suggest that neural tracking at 10 months might not require additional visual bottom-up cues in speech, at least in ideal listening conditions. Yet, it is possible that the lack of a facilitatory effect of visual speech cues might only hold for phase tracking of infant-directed speech, but not amplitude, since phase and amplitude tracking are argued to follow different developmental trajectories (Ortiz Barajas et al., 2021). Therefore, whether and to what extent visual speech cues affect infants' amplitude tracking remains to be seen.

Real-life listening conditions are usually much noisier than those in the current study, in which one speaker produced infant-directed speech in a clear and engaging manner. In fact, speech directed to infants usually co-occurs with background noise and other speakers' utterances (Barker & Newman, 2004; Erickson & Newman, 2017). Previous research with infants and adults has suggested an AV speech benefit especially in noisy and challenging conditions, such as multi-speaker interference (Lalonde & Werner, 2019; Moradi et al., 2013; Sumby & Pollack, 1954). Investigating adults' neural tracking of speech, Zion Golumbic, Cogan, et al. (2013) demonstrated a facilitatory effect of visual speech cues only in a multi-speaker (but not a single-speaker) environment. Therefore, it is possible that visual speech cues aid infants' neural tracking of speech in more challenging conditions, but under ideal listening conditions, such cues are not required to reach a sufficient level of neural tracking. Given that successful perception of speech-in-noise develops through childhood and continues to mature until adulthood (Bertels et al., 2023), longitudinal studies examining the contribution of visual speech cues to speech-in-noise tracking are required.

Another consideration is the difference in study designs and analysis techniques used in different studies. Regarding the study design, whereas previous studies compared AV speech to AO and VO speech, using still-face pictures of the speaker's face for the AO condition, only the mouth region was specifically occluded with a block in the current study, leaving the other visual cues such as the naturalistic movements of the eyes, eyebrows, and the head intact. While visual cues that come from the movement of the mouth and lips are argued to provide the most salient temporal cues to auditory speech given the close temporal correspondence between the articulatory movements and syllable onsets (Chandrasekaran et al., 2009), other cues such as head and eye(brow) movements are also linked to speech perception by conveying information about the speech amplitude envelope, especially with regards to speech prosody, indicating for example phrase boundaries (Cavé et al., 1996; de la Cruz-Pavía et al., 2020; Kim et al., 2014; Munhall et al., 2004). Therefore, it is difficult to conclude whether the AV-benefit observed in several previous studies is truly linked to articulatory cues, or whether it is the facial cues as a whole that allow listeners to form predictions about the speech envelope, thus facilitating tracking. This explanation is also in line with the findings of Haider et al. (2023) who demonstrated that tracking of lip movements specifically further increased adults' tracking of the speech spectrogram only in the single-speaker condition (but not multi-speaker), whereas tracking was generally higher in the AV condition compared to A-only (also when lip movements were blocked with a face mask in the AV condition). They suggested that this enhancement might therefore not be specific to lip movements, but might rather be linked to visual speech in general, indicating a possible role of other facial cues in visual speech processing (Haider et al., 2023). Further research is needed to test how articulatory cues and other facial cues separately contribute to neural tracking of speech.

A final possible explanation for the difference between these results and those of some of the previous studies concerns the metric used to assess neural tracking. While this study employed coherence (similar to Zion Golumbic, Cogan, et al., 2013 who used inter-trial coherence), others, such as Tan et al. (2022), used multivariate temporal response functions (mTRFs) as their neural tracking measure (also see Crosse et al., 2015). One difference between coherence and TRF measures is that coherence deals with the degree of phase consistency between the two signals, but not the magnitude of the phase shift, whereas TRFs provide temporal information on the relationship between the signals, which can go beyond the "static" coherence measure (Chen et al., 2023). As TRFs provide a different metric to quantify the relationship between the neural data and the amplitude envelope of AV speech, TRFs might, for instance, measure a possible difference in the phase shift values between the two conditions, thus yielding more subtle information in the temporal dimension. Although some TRF measures were found to be highly correlated with speech-brain coherence measures in adult-directed speech (Chen et al., 2023), the two measures have

not been compared with regard to infant-directed or AV speech, which remains an important issue for future research.

Regarding language development, we predicted that infants' low-frequency neural tracking of speech would relate to later language outcomes. Indeed, we observed a significant relationship between neural tracking and later vocabulary size. More specifically, neural tracking at the syllable rate predicted infants' subsequent receptive vocabulary, controlling for receptive vocabulary size at the time of the EEG recording. Additionally, stress-rate tracking was related to expressive vocabulary outcomes, and infants with better neural tracking abilities produced more words at 18 months. Our results are in line with recent studies indicating positive links between early neural tracking abilities and subsequent parent-reported language outcomes, which have been demonstrated for stress-rate tracking (Menn, Ward, et al., 2022; Ní Choisdealbha et al., 2023, 2024), syllable-rate tracking (Çetinçelik et al., 2023), and broadly tracking the delta band (~0.5–4 Hz) rhythm (Attaheri, Ní Choisdealbha, Rocha, et al., 2022). This predictive relationship might be related to infants' word segmentation abilities. Tracking the stress and syllable-rate information might highlight the important cues in speech, such as syllables and stressed syllables (Goswami, 2019). Thus, infants who are more adept at neural tracking of stress and syllable rates may have enhanced sensitivity to acoustic cues in speech, which can serve as cues for word onset, and therefore can aid infants' word segmentation (Jusczyk et al., 1999). Early word segmentation abilities have been linked to successful vocabulary outcomes, for both receptive and expressive vocabulary (Junge et al., 2012; Kidd et al., 2018; Kooijman et al., 2013). While our results support previous findings suggesting that infants' tracking of low-frequency information in continuous speech may be relevant for building a robust vocabulary inventory, whether stress and syllable-rate tracking have unique contributions to language development, as tentatively suggested by the current results, remains a question for future research, as well as the possible mediating role of word segmentation. It should be noted that the current data should be interpreted with caution as the regression analyses may be underpowered; the final sample for the vocabulary analysis consisted of 27 infants, whereas a power analysis (based on Çetinçelik et al., 2023) suggested that a sample size of 41 is required.

Our results also have important and timely implications for infants' language acquisition in times of the global COVID-19 pandemic. For infants growing up during the pandemic, their early interactions typically included speakers with face masks on outside of their households, occluding the visual speech cues. Our results indicate that having restricted visual access to such cues does not necessarily impair successful tracking of the speech signal, even though the infants paid less attention to the speaker's face. This conclusion is consistent with previous studies showing that infants' word segmentation (Singh et al., 2021) and children's speech processing abilities (Schwarz et al., 2022) were not negatively affected by the speaker's use of a face mask. On the other hand, they contradict those of Frota et al. (2022), who compared the word segmentation abilities of infants between 7 and 9 months tested during the pandemic to a pre-COVID study (Butler & Frota, 2018) and reported pre- and post-COVID differences in infants' ability to segment words placed at the utterance-edge. However, the authors also reported that they did not find any direct relationship between infants' daily exposure to masks and their word segmentation abilities and concluded that this might be a multifactorial effect that is not necessarily restricted to the use of face masks. Much more work is needed in this area to evaluate the impact of COVID-19 on early language development.

5 | CONCLUSION

This study investigated the impact of visual speech cues, which are the articulatory movements of the speaker's lips, mouth, and jaw, on infants' neural tracking of continuous AV speech. Our results show that 10-month-old infants display neural tracking of naturalistic AV speech at the stressed syllable and syllable frequencies, which are especially pronounced in infant-directed speech. Infants tracked the rhythmic units both when the speaker's visual speech cues were present and when they were not, even though infants paid greater attention to the speaker's face when the visual speech cues were visible. Moreover, we found that neural tracking of the stress- and syllable-rate information in speech at 10 months predicted infants' vocabulary comprehension and production at 18 months, indicating a link between individual differences in neural tracking and subsequent language development. These results suggest that infants' neural tracking of speech remains robust even when visual speech cues are obscured, highlighting the potential role of neural tracking of speech in facilitating language acquisition.

AUTHOR CONTRIBUTIONS

Melis Çetinçelik: Conceptualisation (equal); data curation (equal); formal analysis (lead); investigation (equal); methodology (equal); project administration (equal); software (equal); supervision (equal); visualisation (equal); writing—original draft (lead); writing—review and

editing (equal). **Antonia Jordan-Barros:** Conceptualisation (equal); data curation (equal); methodology (equal); writing—review and editing (equal). **Caroline F. Rowland:** Conceptualisation (equal); funding acquisition (lead); methodology (equal); project administration (equal); resources (equal); supervision (equal); writing—review and editing (equal). **Tineke M. Snijders:** Conceptualisation (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); writing—review and editing (equal).

ACKNOWLEDGEMENTS

We thank all infants and their caregivers for their participation in this study, and the research assistants Inge van Dijke, Ciske Jansen, Daphne Jansen, Jefta Lagerwerf, Iris Schmits, Inge Stok, and Sam Theunissen for their help with the stimulus materials and testing. This work was funded by the Max Planck Society. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.16492.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on the Open Science Framework, at https://osf.io/buxty/?view_only=fe96b4fafd6e495286ac15f1cf257b1a.

ETHICS APPROVAL STATEMENT

This study was approved by the Ethics Committee of the Faculty of Social Sciences, Radboud University (ECSW-2020-025). The study conforms with the Declaration of Helsinki, and written informed consent was obtained prior to the experiment.

ORCID

Melis Çetinçelik https://orcid.org/0000-0002-8931-5732

REFERENCES

Attaheri, A., Ní Choisdealbha, Á., Di Liberto, G. M., Rocha, S., Brusini, P., Mead, N., Olawole-Scott, H., Boutris, P., Gibbon, S., Williams, I., Grey, C., Flanagan, S., & Goswami, U. (2022). Delta- and theta-band cortical tracking and phase-amplitude coupling to sung speech by infants. *NeuroImage*, *247*, 118698. https://doi.org/10.1016/j.neuroimage.2021.118698

Attaheri, A., Ní Choisdealbha, Á., Rocha, S., Brusini, P., Liberto, G. M. D., Mead, N., Olawole-Scott, H., Boutris, P., Gibbon, S., Williams, I., Grey, C., Oliveira, M. A.e, Brough, C., Flanagan, S., & Goswami, U. (2022). *Infant low-frequency EEG cortical power, cortical tracking and phase-amplitude coupling predicts language a year later* (p. 2022.11.02.514963). bioRxiv. https://doi.org/10.1101/2022.11.02.514963

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2* [dataset]. Linguistic Data Consortium. https://doi.org/10.35111/GS6S-GM48

Barker, B. A., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition*, *94*(2), B45–B53. https://doi.org/10.1016/j.cognition.2004.06.001

Bastianello, T., Keren-Portnoy, T., Majorano, M., & Vihman, M. (2022). Infant looking preferences towards dynamic faces: A systematic review. *Infant Behavior and Development*, *67*, 101709. https://doi.org/10.1016/j.infbeh.2022.101709

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. https://www.frontiersin.org/articles/10.3389/fnsys.2015.00175

Bauer, A.-K. R., Debener, S., & Nobre, A. C. (2020). Synchronisation of neural oscillations and cross-modal influences. *Trends in Cognitive Sciences*, *24*(6), 481–495. https://doi.org/10.1016/j.tics.2020.03.003

Begus, K., & Bonawitz, E. (2020). The rhythm of learning: Theta oscillations as an index of active learning in infancy. *Developmental Cognitive Neuroscience*, *45*, 100810. https://doi.org/10.1016/j.dcn.2020.100810

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1–4), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252. https://doi.org/10.3758/BF03205546

Bertels, J., Niesen, M., Destoky, F., Coolen, T., Vander Ghinst, M., Wens, V., Rovai, A., Trotta, N., Baart, M., Molinaro, N., De Tiège, X., & Bourguignon, M. (2023). Neurodevelopmental oscillatory basis of speech processing in noise. *Developmental Cognitive Neuroscience*, *59*, 101181. https://doi.org/10.1016/j.dcn.2022.101181

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. https://doi.org/10.1016/j.crneur.2021.100014

Blackmagic Design. (2020). DaVinci Resolve (Version 16) [Software]. *Blackmagic Design*. https://www.blackmagicdesign.com/products/davinciresolve

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [Computer program]. http://www.praat.org/

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *The Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

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

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

Cantiani, C., Dondena, C., Molteni, M., Riva, V., & Piazza, C. (2022). Synchronizing with the rhythm: Infant neural entrainment to complex musical and speech stimuli. *Frontiers in Psychology*, *13*, 944670. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.944670

Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and Fo variations. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, 4, 2175–2178. https://doi.org/10.1109/ICSLP.1996.607235

Çetinçelik, M., Rowland, C. F., & Snijders, T. M. (2023). Ten-month-old infants' neural tracking of naturalistic speech is not facilitated by the speaker's eye gaze. *Developmental Cognitive Neuroscience*, *64*, 101297. https://doi.org/10.1016/j.dcn.2023.101297

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Chen, Y.-P., Schmidt, F., Keitel, A., Rösch, S., Hauswald, A., & Weisz, N. (2023). Speech intelligibility changes the temporal evolution of neural speech tracking. *NeuroImage*, *268*, 119894. https://doi.org/10.1016/j.neuroimage.2023.119894

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. 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. *The Journal of Neuroscience*, *36*(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

de la Cruz-Pavía, I., Gervain, J., Vatikiotis-Bateson, E., & Werker, J. F. (2020). Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech. *Language Acquisition*, *27*(2), 160–186. https://doi.org/10.1080/10489223.2019.1659276

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

Erickson, L. C., & Newman, R. S. (2017). Influences of background noise on infants and children. *Current Directions in Psychological Science*, *26*(5), 451–457. https://doi.org/10.1177/0963721417709087

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, *53*(1), 18–33. https://doi.org/10.1044/1092-4388(2009/08-0140)

Frota, S., Pejovic, J., Cruz, M., Severino, C., & Vigário, M. (2022). Early word segmentation behind the mask. *Frontiers in Psychology*, *13*, 879123. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.879123

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Goswami, U. (2019). Speech rhythm and language acquisition: An amplitude modulation phase hierarchy perspective. *Annals of the new York Academy of Sciences*, *1453*(1), 67–78. https://doi.org/10.1111/nyas.14137

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

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1275–1281. https://doi.org/10.1037/xhp0000208

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2023). Neural speech tracking highlights the importance of visual speech in multi-speaker situations. *Journal of Cognitive Neuroscience*, *1–15*, 128–142. https://doi.org/10.1162/jocn_a_02059

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. https://doi.org/10.1111/j.1467-8624.2005.00866.x

Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. (2011). Neural signatures of face–voice synchrony in 5-month-old human infants. *Developmental Psychobiology*, *53*(4), 359–370. https://doi.org/10.1002/dev.20525

Jessen, S., Fiedler, L., Münte, T. F., & Obleser, J. (2019). Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie. *NeuroImage*, *202*, 116060. https://doi.org/10.1016/j.neuroimage.2019.116060

Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development: Rapid recognition. *Developmental Science*, *15*(4), 463–473. https://doi.org/10.1111/j.1467-7687.2012.1144.x

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

Kalashnikova, M., Peter, V., Liberto, G. M. D., Lalor, E. C., & Burnham, D. (2018). Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific Reports*, *8*, 13745. https://doi.org/10.1038/s41598-018-32150-6

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in infant speech segmentation: Achieving the lexical shift. *Infancy*, *23*, 770–794. https://doi.org/10.1111/infa.12256

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, *57*, 317–330. https://doi.org/10.1016/j.specom.2013.06.003

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

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

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105

Kushnerenko, E., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., & Moore, D. (2013). Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. *Frontiers in Psychology*, *4*, 432. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00432

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

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

Lalonde, K., & McCreery, R. W. (2020). Audiovisual enhancement of speech perception in noise by school-age children who are hard of hearing. *Ear and Hearing*, *41*(4), 705–719. https://doi.org/10.1097/AUD.0000000000000830

Lalonde, K., & Werner, L. A. (2019). Infants and adults use visual cues to improve detection and discrimination of speech in noise. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3860–3875. https://doi.org/10.1044/2019_JSLHR-H-19-0106

Lalonde, K., & Werner, L. A. (2021). Development of the mechanisms underlying audiovisual speech perception benefit. *Brain Sciences*, *11*(1), 49. https://doi.org/10.3390/brainsci11010049

Leong, V., Kalashnikova, M., Burnham, D., & Goswami, U. (2017). The temporal modulation structure of infant-directed speech. *Open Mind*, *1*(2), 78–90. https://doi.org/10.1162/opmi_a_00008

Leong, V., & Goswami, U. (2015). Acoustic-emergent phonology in the amplitude envelope of child-directed speech. *PLoS ONE*, *10*(12), e0144411. https://doi.org/10.1371/journal.pone.0144411

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*, 66–77. https://doi.org/10.1037/a0015579

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

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. 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. https://doi.org/10.1371/journal.pbio.1000445

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

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

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* (Vol. 8, pp. 145–151). MIT Press.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. 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. https://doi.org/10.1038/264746a0

Menn, K. H., Michel, C., Meyer, L., Hoehl, S., & Männel, C. (2022). Natural infant-directed speech facilitates neural tracking of prosody. *NeuroImage*, *251*, 118991. https://doi.org/10.1016/j.neuroimage.2022.118991

Menn, K. H., Ward, E. K., Braukmann, R., van den Boomen, C., Buitelaar, J., Hunnius, S., & Snijders, T. M. (2022). Neural tracking in infancy predicts language development in children with and without family history of autism. *Neurobiology of Language*, *3*(3), 495–514. https://doi.org/10.1162/nol_a_00074

Mishra, S., Stenfelt, S., Lunner, T., Rönnberg, J., & Rudner, M. (2014). Cognitive spare capacity in older adults with hearing loss. *Frontiers in Aging Neuroscience*, *6*, 96. https://www.frontiersin.org/articles/10.3389/fnagi.2014.00096

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

Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, *55*(8), 1640–1655. https://doi.org/10.1037/dev0000750

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

Ní Choisdealbha, Á., Attaheri, A., Rocha, S., Mead, N., Olawole-Scott, H., Brusini, P., Gibbon, S., Boutris, P., Grey, C., Hines, D., Williams, I., Flanagan, S. A., & Goswami, U. (2023). Neural phase angle from two months when tracking speech and non-speech rhythm linked to language performance from 12 to 24 months. *Brain and Language*, *243*, 105301. https://doi.org/10.1016/j.bandl.2023.105301

Ní Choisdealbha, Á., Attaheri, A., Rocha, S., Mead, N., Olawole-Scott, H., Alfaro e Oliveira, M., Brough, C., Brusini, P., Gibbon, S., Boutris, P., Grey, C., Williams, I., Flanagan, S., & Goswami, U. (2024). Cortical tracking of visual rhythmic speech by 5- and 8-month-old infants: Individual differences in phase angle relate to language outcomes up to 2 years. *Developmental Science*, *27*(4). https://doi.org/10.1111/desc.13502

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of MEG,

EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. 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. https://doi.org/10.1016/j.clinph.2005.12.027

Ortiz Barajas, M. C., Guevara, R., & Gervain, J. (2021). The origins and development of speech envelope tracking during the first months of life. *Developmental Cognitive Neuroscience*, *48*, 100915. https://doi.org/10.1016/j.dcn.2021.100915

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. 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. 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. https://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

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

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

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

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. https://doi.org/10.3389/fpsyg.2012.00216

R Core Team. (2022). *R: A language and environment for statistical computing (4.2.2) [computer software]*. R Foundation for Statistical Computing. http://www.R-project.org/

Reynolds, G. D., Bahrick, L. E., Lickliter, R., & Guy, M. W. (2014). Neural correlates of intersensory processing in 5-month-old infants. *Developmental Psychobiology*, *56*(3), 355–372. https://doi.org/10.1002/dev.21104

Rosenberg, J. R., Amjad, A. M., Breeze, P., Brillinger, D. R., & Halliday, D. M. (1989). The fourier approach to the identification of functional coupling between neuronal spike trains. *Progress in Biophysics and Molecular Biology*, *53*(1), 1–31. https://doi.org/10.1016/0079-6107(89)90004-7

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *The European Journal of Neuroscience*, *33*(12), 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

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

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

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

Schwartz, J.-L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*, *10*(7), e1003743. https://doi.org/10.1371/journal.pcbi.1003743

Schwarz, J., Li, K. K., Sim, J. H., Zhang, Y., Buchanan-Worster, E., Post, B., Gibson, J. L., & McDougall, K. (2022). Semantic cues modulate children's and adults' processing of audio-visual face mask speech. *Frontiers in Psychology*, *13*, 879156. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.879156

Sekiyama, K., Hisanaga, S., & Mugitani, R. (2021). Selective attention to the mouth of a talker in Japanese-learning infants and toddlers: Its relationship with vocabulary and compensation for noise. *Cortex*, *140*, 145–156. https://doi.org/10.1016/j.cortex.2021.03.023

Singh, L., Tan, A., & Quinn, P. C. (2021). Infants recognize words spoken through opaque masks but not through clear masks. *Developmental Science*, *24*(6), e13117. https://doi.org/10.1111/desc.13117

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. https://doi.org/10.1121/1.1907309

Tan, S. H. J., Kalashnikova, M., Di Liberto, G. M., Crosse, M. J., & Burnham, D. (2022). Seeing a talking face matters: The relationship between cortical tracking of continuous auditory-visual speech and gaze behaviour in infants, children and adults. *NeuroImage*, *256*, 119217. https://doi.org/10.1016/j.neuroimage.2022.119217

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. https://doi.org/10.1016/j.cognition.2008.05.009

Thézé, R., Giraud, A.-L., & Mégevand, P. (2020). The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. *Science Advances*, *6*(45), eabc6348. https://doi.org/10.1126/sciadv.abc6348

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in school-age children: The roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research*, *57*(2), 556–565. https://doi.org/10.1044/2013_JSLHR-H-12-0273

Vanden Bosch der Nederlanden, C. M., Joanisse, M. F., Grahn, J. A., Snijders, T. M., & Schoffelen, J.-M. (2022). Familiarity modulates neural tracking of sung and spoken utterances. *NeuroImage*, *252*, 119049. https://doi.org/10.1016/j.neuroimage.2022.119049

Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory integration as a window into orderly and disrupted cognition and communication. *Annual Review of Psychology*, *71*(1), 193–219. https://doi.org/10.1146/annurev-psych-010419-051112

Watson, T. L., Robbins, R. A., & Best, C. T. (2014). Infant perceptual development for faces and spoken words: An integrated approach. *Developmental Psychobiology*, *56*(7), 1454–1481. https://doi.org/10.1002/dev.21243

Zink, I., & Lejaegere, M. (2003). *N-CDIs: Korte vormen: Lijsten voor communicatieve ontwikkeling: Aanpassing en hernormering van de MacArthur short form Vocabulatory checklists van Fenson et al*. ACCO.

Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, *33*(4), 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zion Golumbic, E., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zoefel, B. (2021). Visual speech cues recruit neural oscillations to optimise auditory perception: Ways forward for research on human communication. *Current Research in Neurobiology*, *2*, 100015. https://doi.org/10.1016/j.crneur.2021.100015

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Çetinçelik, M., Jordan-Barros, A., Rowland, C. F., & Snijders, T. M. (2024). The effect of visual speech cues on neural tracking of speech in 10-month-old infants. *European Journal of Neuroscience*, *60*(6), 5381–5399. https://doi.org/10.1111/ejn.16492