Four-Year-Olds' Cortical Tracking to Continuous Auditory-Visual Speech

S.H. Jessica Tan[1], *Michael J. Crosse*[2], *Giovanni M. Di Liberto*[3], *Denis Burnham*[1]

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

[2]The Sheryl and Daniel R. Tishman Cognitive Neurophysiology Laboratory, Department of Pediatrics, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY 10461, USA

[3]Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL University, France

j.tan@westernsydney.edu.au, michael.crosse@einstein.yu.edu, diliberg@tcd.ie
denis.burnham@westernsydney.edu.au

Abstract

Visual speech information, such as a speaker's mouth and eyebrow movements, enhances speech perception. Evidence for this perceptual benefit has mainly been from behavioural or neurophysiological studies that made use of event-related potentials (ERPs). ERP studies, however, are limited by repetitive and short stimuli that are not representative of natural speech. An approach that examines cortical tracking of the speech envelope allows for the use of continuous speech stimuli. This approach has recently been employed to demonstrate that adults' cortical tracking of the speech envelope is augmented when synchronous visual speech information is provided [1]. To date, no study has investigated whether children, like adults, show stronger envelope tracking when congruent visual speech information is available. This study investigates this question by measuring four-year-olds' cortical tracking of continuous auditory-visual speech through electroencephalography (EEG). Cortical tracking was quantified by means of ridge regression models that estimate the linear mapping from the speech to the EEG signal and vice versa. Stimulus reconstruction for auditory-only and auditory-visual speech was found to be stronger compared to visual-only speech.

Index Terms: cortical tracking, auditory-visual speech perception, visual speech benefit

1. Introduction

Research on auditory-visual speech perception has established that visual speech information, such as a speaker's lip and head movements, contributes to and augments speech perception and comprehension. The perceptual benefit resulting from the addition of visual speech information to auditory information is known as visual speech benefit (VSB). Behavioural studies have found that VSB is present across development (children: [2]; adults: [3]), but that it increases with age [4]. Support for these behavioural findings come from neurophysiological studies that have mainly investigated event-related potentials (ERPs). Specifically, congruent visual speech input was shown to reduce amplitudes and shortened latencies of auditory N1 and P2 components were found in children (e.g., [5]) and adults (e.g., [6]) when auditory-visual stimuli compared to auditory-only stimuli were presented. However, the ERP approach usually requires the use of repetitive and short stimuli, such as isolated syllables or words that are not representative of natural continuous speech.

Recent research has demonstrated solutions to investigate speech perception with naturalistic speech stimuli. One such approach involves assessing the coupling between brain signals and a continuous sensory input. Specifically, the synchronization between the temporal envelope of a speech input and the corresponding brain responses, a measure that is referred to as cortical tracking of the speech envelope, or more simply as *envelope tracking*. Even though this approach has been increasingly used to examine auditory-only speech perception in adults (e.g., [7]-[8]), relatively little is known about how reliably the cortical signals track visual speech information. The few studies conducted so far suggest that seeing a speaker's talking face augments envelope tracking [1],[9]-[10]. Importantly, although there is evidence that cortical tracking of auditory speech can be measured in children [11]-[13], there are no studies available in the literature on auditory-visual speech perception in children to our knowledge.

This study addresses this gap by investigating cortical tracking of auditory-visual speech in children. As in [1], cortical tracking was indexed by means of ridge regression models describing the mapping between the speech envelope and the EEG signal. The regression fit was then used to predict the EEG (forward modelling) and reconstruct the speech envelope (backward modelling) in unseen data with cross-validation, and the quality of those predictions was taken as a measure of envelope tracking. Stronger envelope tracking was expected when participants were presented with auditory-visual speech compared to auditory-only and visual-only speech conditions.

2. Method

2.1. Participants

Fourteen 4-year-old Australian-English monolinguals were recruited. These children were born full-term, with normal hearing and vision, no known history of ear infections and cognitive or language delay.

2.2. Stimuli

Auditory-visual recordings of 30 utterances (e.g., "*How are you today? You look happy! Are you ready for some fun?")* were made in infant-directed speech by a female native speaker of Australian English. The recordings consisted of a close-up of the speaker's face and shoulders against a white background.

There were three presentation modes that made up the conditions: auditory-only (AO), visual-only (VO) and auditory-visual (AV). Auditory-only and visual-only recordings were extracted separately from the AV recordings.

In the AO condition, a still image of the speaker's resting face was shown on the screen as the auditory track played. In the VO condition, the dynamic video of the speaker's talking face was presented in silence. In the AV condition, the dynamic video and its soundtrack were presented synchronously. The 30 utterances were presented in three blocks. Each block consisted of 10 utterances that were presented once in each modality (10 x 3 = 30 trials). Presentation order of trials were randomized across modalities in such a manner that no two modalities of the same utterance appeared side by side. A 3-s cartoon animation was played at the end of each block. Attention-getters consisted of different pictures of characters from the *Minions* movie that appeared in a randomly after either two or three trials.

2.3. Procedure

Participants sat on a chair approximately 70cm away from the centre of a 17-inch DELL LCD monitor which played the video recordings while auditory recordings were played via two loudspeakers (Edirol MA-15 Digital Stereo Micro Monitors) placed at the left and right sides of the monitor. Continuous EEG data were recorded with a 128-channel Hydrocel Geodesic Sensor Net (HCGSN), NetAmps 300 amplifier, and NetStation 4.5.7 software (EGI Inc) at a sampling rate of 1000Hz, with the reference electrode placed at Cz, saved for offline analyses. Stimulus presentation was controlled via Presentation software (Neurobehavioural Systems). Eye-tracking recordings were co-registered with EEG recordings through a Tobii extension in the Presentation software. A Tobii X120 eye-tracker was placed below the LCD screen to gather gaze fixation data.

To motivate the child participants to focus on the screen, the session was framed as a game in which they were required to press a button on a response pad whenever they spotted a picture of a character from the *Minions* movie [5].

2.4. Pre-processing

2.4.1. EEG data

EEG data were pre-processed using EEGLAB, FieldTrip, NoiseTools, and custom scripts in MATLAB R2018b. First, EEG data from the three outer rings of the net were removed because these channels have been found to be very noisy in children because they strongly reflected motor movements [14]. EEG data from the remaining 92 channels were filtered using *pop_eegfiltnew()* in EEGLAB with 1Hz as the high-pass cut-off and 8Hz as the low-pass cut-off, down-sampled to 200Hz, and re-referenced to the average of all remaining channels. Artifact subspace reconstruction (ASR) was applied to remove noise. A sliding window of 500ms and threshold of 5 standard deviations were used to identify corrupted subspaces. The noisy channels that were removed during ASR were replaced with an estimate of the neighbouring clean channels via spherical spline interpolation (EEGLAB; [15]).

2.4.2. Stimuli

Stimuli were presented at 48kHz but filtered and down-sampled to 200Hz to match the sampling rate of the EEG data and characterized using the broadband speech envelope of the acoustic signal. A spectrogram representation of each stimulus was generated using a compressive gammachirp auditory filter bank that modelled the auditory periphery [16]. The envelope at each of the 128 frequency bands was calculated using a Hilbert transform and the broadband envelope was obtained by averaging across the 128 narrow-band envelopes.

2.5. Data Analyses

EEG data analyses were conducted using the mTRF toolbox [17], PERMUTOOLS, and custom scripts in MATLAB R2018b.

2.5.1. Temporal response functions (TRFs)

Temporal response functions (TRFs) were used to estimate the linear function describing the mapping between neural responses and the speech envelope at every channel and can be interpreted in terms of their spatio-temporal dynamics. Envelope tracking in all three conditions was indexed by relative TRF fit and the accuracy of production of the unseen EEG signal based on the envelope of the speech signal. The TRFs obtained for the three conditions were used to predict the EEG signal using leave-one-out cross-validation. Pearson's correlation values between the recorded and predicted EEG signal were used to index sensor-space neural entrainment for each participant, resulting in a distribution of r values for each participant and electrode.

Permutation analyses with false discovery rate (FDR) correction were conducted to (1) detect any difference between TRFs for AO, VO, and AV conditions, and (2) identify whether any group of electrodes consistently tracked the speech envelope.

2.5.2. Stimulus reconstruction

The stimulus reconstruction method [1][9] was also used to examine neural entrainment. This involves the identification of linear decoders that describe the optimal linear mapping from the EEG data to the speech envelope of the stimulus. Envelope reconstruction accuracy was quantified by means of a Pearson's correlation between the estimated and the original speech envelopes. This produced a distribution of r values for each participant and trial.

Linear mixed-effects modelling analyses were conducted to establish the differences, if any, between AO, VO and AV conditions.

3. Results

3.1. TRFs

To assess if there was any difference between TRFs for the three conditions during the EEG recording, a permutation analysis with FDR correction was employed. The test did not reveal any significant difference between conditions on any electrode or time point. Similar analyses that were conducted with groups of electrodes representing the frontocentral and occipital regions did not reveal any significant difference between conditions as well ($p > .12$). Figures 1 and 2 depict the temporal response functions at the frontocentral region (Fig. 1) and the occipital scalp regions (Fig. 2). Figure 3 shows the r values that describe the correlation between the recorded and predicted EEG signal.

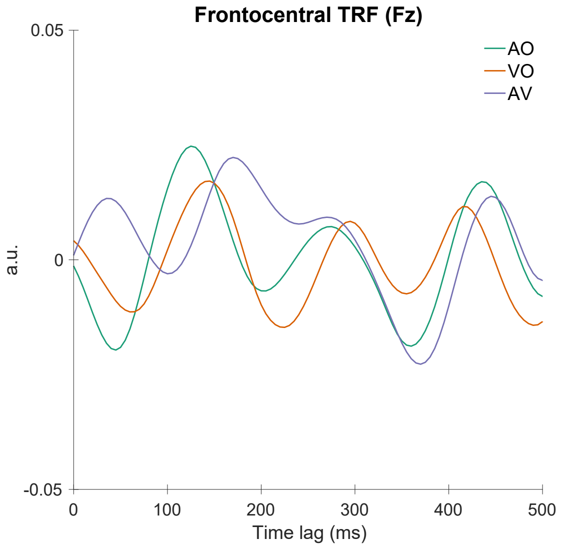

Figure 1. *Group-average TRF at the frontocentral channels.*

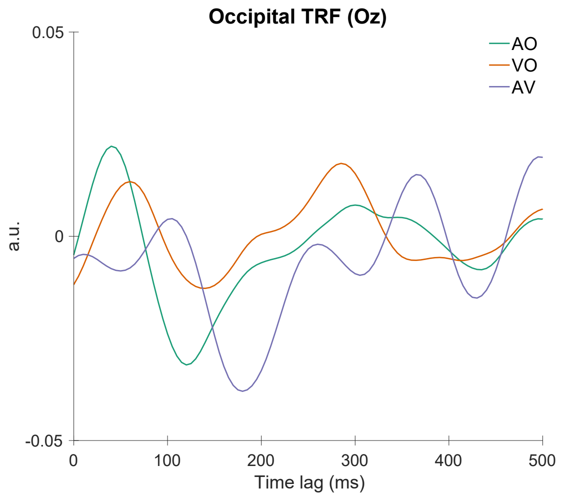

Figure 2. *Group-average TRF at the occipital channels.*

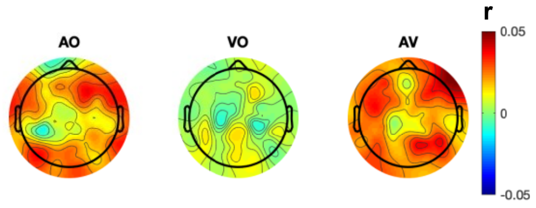

Figure 3. *Scalp topographies of the EEG predictions.*

A measure of global field power (GFP) was estimated by calculating the standard deviation of the TRFs across all 92 channels (Figure 4). GFP constitutes a reference-independent measure of response strength across the entire scalp at each time lag [18]. The temporal profile of the GFP measure suggests that two clear TRF components are evident for AO (~50ms and ~130ms) and AV (~35ms and ~175ms). Therefore, paired-samples t-tests were conducted to investigate whether there is any difference between conditions. None of the tests revealed any significant difference between conditions (all *ps* > .17).

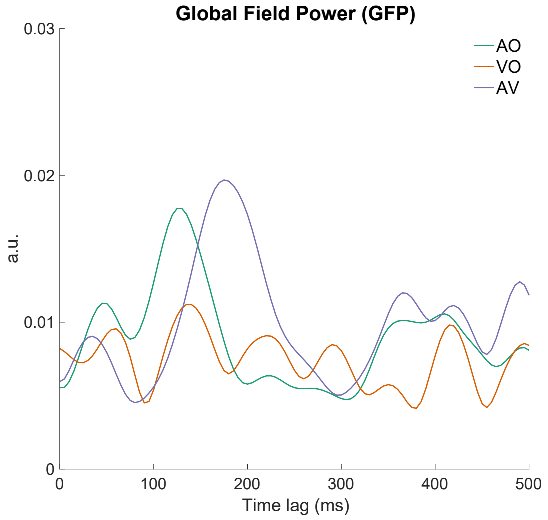

Figure 4. *Global field power (GFP) measured at each time lag.*

3.2. Stimulus reconstruction

Envelope reconstructions were derived by means of backward TRF models. The quality of that linear mapping, which was quantified by means of Pearson's correlation between the actual envelope and its reconstruction, was considered as a measure of envelope tracking. A Linear Mixed Effect Regression (LMER) model with Pearson's *r* as the dependent variable, condition as a fixed factor, and subject and trials completed as random intercepts.

Only the main effect of condition was significant, *F*(2, 759.11) = 24.44, *p* < .001. The Kenward-Roger approximation to the degrees of freedom was used to calculate the *p*-values for the fixed effect of condition [19], and the *ANOVA* function from the *car* package in R with test specified as "F" were used. Next, to further examine the effect of condition, multiple comparisons were conducted via the R-package *lsmeans*. There were significant differences between AO and VO conditions (*t*(759) = 6.25, *p* < .0001), and between AV and VO conditions (*t*(759) = 5.85, *p* < .0001). The difference between AO and AV conditions was not significant (*t*(759) = 0.40, *p* = .92). Stimulus reconstruction accuracy was significantly higher for AO (*M* = 0.08, *SE* = 0.008) and AV (*M* = 0.08, *SE* = 0.008) conditions compared to VO (*M* = 0.02, *SE* = 0.008) condition but was similar for AO and AV conditions.

4. Discussion

This study investigated whether four-year-olds' cortical tracking of the continuous speech envelope is enhanced by visual speech information by comparing envelope tracking when participants were presented with auditory-only, visual-only, and auditory-visual speech inputs. Cortical tracking was examined via temporal response functions and stimulus reconstruction.

When cortical tracking was indexed by forward TRF models, no significant difference was found between conditions at any channel or time point. When cortical tracking was indexed by envelope reconstruction correlations, AO and AV conditions had significantly greater reconstruction accuracy than the VO condition. Contrary to our hypotheses, AO and AV conditions did not show any significant differences. These findings are surprising given that behavioural studies have shown that children benefit from visual speech information on speech perception tasks (e.g., [2]). However, it is possible that the effect of visual information could not be measured because it was too weak for the particular set of stimuli used in this experiment which consisted of infant-directed speech that had facilitative prosodic cues without noise nor competing talkers or sounds. Another possible explanation, as suggested by findings from [20], may be that 4-year-olds prefer auditory information when processing multisensory events, and may therefore not be placing as much importance on visual information. Further analyses and studies are necessary in order to obtain a clearer understanding of this issue.

Preliminary analyses indicate stronger stimulus reconstruction of auditory-only and auditory-visual speech compared to visual-only speech. Further investigations need to be conducted to examine the differences between findings from the temporal response function approach and the stimulus reconstruction approach. Another step for future work is to establish whether individual differences in gaze behavior modulate the strength of cortical tracking of the auditory-visual speech envelope.

5. Acknowledgements

This research was supported by a doctoral scholarship to the first author funded by the MARCS Institute at Western Sydney University and the HEARing Cooperative Research Centre (CRC), and by HEARing CRC funding to the fourth author. We gratefully appreciate the assistance of Varghese Peter, Caitlin Hooper, the MARCS technical team, and the parents and their children who participated in the study.

6. References

[1] M. J. Crosse, J. S. Butler, and E. C. Lalor, "Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions," *Journal of Neuroscience,* vol. 35, no. 42, pp. 14195-14204, 2015.

[2] L. A. Ross, S. Molholm, D. Blanco, M. Gomez-Ramirez, A. Saint-Amour, and J. J. Foxe, "The development of multisensory speech perception continues into the late childhood years," *European Journal of Neuroscience*, vol. 33, no. 12, pp. 2329-2337, 2011.

[3] S. Moradi, B. Lidestam, and J. Rönnberg, "Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy," *Frontiers in Psychology*, vol. 4, pp. 1-13, 2013.

[4] S. Jerger, M. F. Damian, N. Tye-Murray, and H. Abdi, "Children use visual speech to compensate for non-intact auditory speech," *Journal of Experimental Child Psychology*, vol. 126, pp. 295-312, 2014.

[5] N. Kaganovich, and J. Schumaker, "Audiovisual integration for speech during mid-childhood: Electrophysiological evidence," *Brain and Language,* vol. 139, pp. 36-48, 2014.

[6] M. Baart, and A. G. Samuel, "Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing," *Journal of Memory and Language,* vol. 85, pp. 42-59, 2015.

[7] N. Ding, and J. Z. Simon, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *Journal of Neuroscience*, vol. 33, no. 13, pp. 5728-5735, 2013.

[8] J. M. Rimmele, E. Zion Golumbic, E. Schröger, and D. Poeppel, "The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene," *Cortex*, vol. 68, pp. 144-154, 2014.

[9] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor, "Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," *Journal of Neuroscience*, vol. 26, no. 28, pp. 9888-9895, 2016.

[10] Zion Golumbic E., G. B. Cogan, E. Schröger, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"," *Journal of Neuroscience,* vol. 33, no. 4, pp. 1417-1426, 2013.

[11] G. M. Di Liberto, V. Peter, M. Kalashnikova, U. Goswami, D. Burnham, and E. C. Lalor, "Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia," *Neuroimage,* vol. 175, pp. 70-79, 2018.

[12] M. Kalashnikova, V. Peter, G. M. Di Liberto, E. C. Lalor, and D. Burnham, "Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech," *Nature*, vol. 8., pp. 13745, 2018.

[13] A. J. Power, N. Mead, L. Barnes, and U. Goswami, "Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children," *Frontiers in Psychology*, vol. 3, pp. 216, 2012.

[14] N. A. Folland, B. E. Butler, J. E. Payne, and L. J. Trainor, "Cortical representations sensitive to the number of perceived auditory objects between 2 and 4 months of age: Electrophysiological evidence," *Journal of Cognitive Neuroscience*, vol. 27, no. 5, pp. 1060–1067, 2015.

[15] A. Delorme, and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no.1, pp. 9–21, 2004.

[16] T. Irino, and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Trans Audio Speech Lang Process,* vol. 14, pp. 2222-2232, 2006.

[17] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience,* vol. 10, pp. 3958-14, 2016.

[18] D. Lehmann and W. Skrandies, "Reference-free identification of components of checkerboard-evoked multichannel potential fields," *Electroencephalogr. Clin. Neurophysiol,* vol. 48, pp. 609–621, 1980.

[19] U. Halekoh and S. Hojsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," J. Stat. Softw., vol. 59, no. 9, pp. 1–30, 2014.

[20] R. J. Hirst, J. E. Stacey, L. Cragg, P. C. Stacey, and H. A. Allen, "The threshold for the Mcgurk effect in audio-visual noise decreases with development," *Scientific Reports*, 8, pp. 12372, 2018.