




Seeing a Talking Face Matters: Gaze Behavior and the Auditory–Visual Speech Benefit in Adults’ Cortical Tracking of Infant-directed Speech

Sok Hui Jessica Tan^{1,2} , Marina Kalashnikova^{3,4}, Giovanni M. Di Liberto⁵, Michael J. Crosse^{6,7}, and Denis Burnham¹

Abstract

■ In face-to-face conversations, listeners gather visual speech information from a speaker’s talking face that enhances their perception of the incoming auditory speech signal. This auditory–visual (AV) speech benefit is evident even in quiet environments but is stronger in situations that require greater listening effort such as when the speech signal itself deviates from listeners’ expectations. One example is infant-directed speech (IDS) presented to adults. IDS has exaggerated acoustic properties that are easily discriminable from adult-directed speech (ADS). Although IDS is a speech register that adults typically use with infants, no previous neurophysiological study has directly examined whether adult listeners process IDS differently from ADS. To address this, the current study simultaneously recorded EEG and eye-tracking data from adult participants as they were presented with auditory-only (AO), visual-only, and AV

recordings of IDS and ADS. Eye-tracking data were recorded because looking behavior to the speaker’s eyes and mouth modulates the extent of AV speech benefit experienced. Analyses of cortical tracking accuracy revealed that cortical tracking of the speech envelope was significant in AO and AV modalities for IDS and ADS. However, the AV speech benefit [i.e., $AV > (A + V)$] was only present for IDS trials. Gaze behavior analyses indicated differences in looking behavior during IDS and ADS trials. Surprisingly, looking behavior to the speaker’s eyes and mouth was not correlated with cortical tracking accuracy. Additional exploratory analyses indicated that attention to the whole display was negatively correlated with cortical tracking accuracy of AO and visual-only trials in IDS. Our results underscore the nuances involved in the relationship between neurophysiological AV speech benefit and looking behavior. ■

INTRODUCTION

The human face holds a plenitude of information. The eye region generally conveys information about a speaker’s emotions (Buchan, Paré, & Munhall, 2007) and intonation patterns (Lansing & McConkie, 1999) whereas the mouth region typically imparts information about the articulatory (Owens & Blazek, 1985) and acoustic (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009) properties of the auditory speech signal. Such visual cues from the speaker’s face are exploited by listeners to decode

the acoustic speech signal; visual speech information from the speaker’s lips and head movements contribute to speech perception and comprehension. Behavioral and neurophysiological studies show that adults and children consistently perform better on speech perception tasks in auditory–visual (AV) as opposed to auditory-only (AO) conditions (e.g., Burnham et al., 2022; Alsius, Wayne, Paré, & Munhall, 2016; Erdener & Burnham, 2013; Ross et al., 2011; Rosenblum, Johnson, & Saldaña, 1996) and that cortical tracking of the speech envelope is enhanced during AV speech compared with AO speech (e.g., Crosse, Butler, & Lalor, 2015). Results from studies conducted using auditory noise reveal that this AV speech benefit is stronger when greater listening effort is required (e.g., Crosse, Di Liberto, & Lalor, 2016; Sumbly & Pollack, 1954). In the absence of noise, a greater listening effort may be required also in situations when the speech signal itself deviates from listeners’ expectations. Such situations arise when the speaker’s speech register does not fit the social expectations of the communicative situation, such as when adults are addressed with infant-directed speech (IDS) in the absence of an infant or non-native-accented speech in the absence of a nonnative listener. Whether and to

¹The MARCS Institute of Brain, Behaviour and Development, Western Sydney University, Australia, ²Science of Learning in Education Centre, Office of Education Research, National Institute of Education, Nanyang Technological University, Singapore, ³The Basque Center on Cognition, Brain and Language, ⁴IKERBASQUE, Basque Foundation for Science, ⁵ADAPT Centre, School of Computer Science and Statistics, Trinity College Institute of Neuroscience, Trinity College, The University of Dublin, Ireland, ⁶SEGOTIA, Galway, Ireland, ⁷Trinity Center for Biomedical Engineering, Department of Mechanical, Manufacturing & Biomedical Engineering, Trinity College Dublin, Dublin, Ireland

what extent the AV speech benefit is experienced in such situations is the primary aim of this study.

The AV Speech Benefit When Speech Deviates from Expectations

An AV speech benefit can occur when listening effort is low (e.g., Fort et al., 2013), but it is stronger when greater listening effort is required. Listening effort has been commonly manipulated via the use of the speech-in-noise paradigm (e.g., Sumbly & Pollack, 1954). In their seminal study, Sumbly and Pollack (1954) varied the speech-to-noise ratio of auditory stimuli in AO and AV presentations and found that AV gain increases with increasing noise. Later studies support this finding: Speech intelligibility in a cocktail-party scenario has been found to be higher in AV than in AO conditions (Schwartz, Berthommier, & Savariaux, 2004), and the degradation of speech comprehension in the presence of multiple competing voices is reduced when the speaker is visible (Rudmann, McCarley, & Kramer, 2003). AV presentations of consonants, words, and sentences in noise result in faster RTs and greater accuracy in identification compared with AO presentations (Moradi, Lidestam, & Rönnerberg, 2013). Together, these findings provide evidence for an AV speech benefit in difficult listening situations and suggest that the extent of this benefit is larger when speech processing is more effortful.

Apart from auditory noise, greater listening effort is also required when the speech signal itself deviates from listeners' expectations, such that a mismatch between the incoming auditory signal and the listener's expectations can result in reduced speech intelligibility (e.g., Gordon-Salant, Yeni-Komshian, & Fitzgibbons, 2010; Gordon-Salant, Yeni-Komshian, Fitzgibbons, & Schurman, 2010; Ferguson, Keum, Jongman, & Sereno, 2009). This situation is relatively understudied although it is a rather common occurrence in daily life. One example is exposure to non-native-accented speech. Even when listeners can correctly repeat or transcribe non-native-accented speech, they still report that accented speech is more difficult to understand (Munro & Derwing, 1995), it is processed more slowly (Porretta, Buchanan, & Järvikivi, 2020; Adank, Evans, Stuart-Smith, & Scott, 2009; Floccia, Butler, Goslin, & Ellis, 2009), and it is comprehended less well than native-accented speech (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). Perhaps unsurprisingly, recognition of accented speech improves when visual speech cues accompany the auditory input (e.g., Banks, Gowen, Munro, & Adank, 2015; Kawase, Hannah, & Wang, 2014; Janse & Adank, 2012; Arnold & Hill, 2001).

Another case in which perceptual challenges can occur because of a mismatch between listeners' expectations and the acoustic properties of the auditory speech input is the use of speech registers that are not intended for the listener or the communicative situation. For instance, although non-native-directed speech—the speech register used in interactions with nonnative speakers of a

language—is rated by nonnative listeners as more intelligible than native-directed speech (Bobb et al., 2019), native listeners rate non-native-directed speech as less intelligible compared with another phonetically exaggerated speech register that is more familiar to them (e.g., Lombard speech, used in noisy situations; Hazan, Uther, & Grunlund, 2015) and report more negative emotions when hearing this register compared with native-directed speech (Knoll, Uther, & Costall, 2011; Knoll & Scharrer, 2007; Uther, Knoll, & Burnham, 2007).

In a similar vein, IDS has been shown to enhance speech processing in its intended audience, that is, infants (e.g., Kalashnikova, Peter, Di Liberto, Lalor, & Burnham, 2018; Bosseler, Teinonen, Tervaniemi, & Huottilainen, 2016; Peter, Kalashnikova, Santos, & Burnham, 2016). Compared with adult-directed speech (ADS), IDS is characterized by slower tempo and speech rate (Narayan & McDermott, 2016), simplified grammar and lexicon (Soderstrom, 2007), exaggerated acoustic properties such as heightened pitch (Burnham, Kitamura, & Vollmer-Conna, 2002), greater pitch range (Fernald et al., 1989), and a positive warm affect (Kitamura & Burnham, 2003). The exaggerated acoustic pitch (Kitamura, Thanavishuth, Burnham, & Luksaneeyanawin, 2001) and prosody (Fernald & Mazzie, 1991) are often paired with exaggerated facial expressions (Chong, Werker, Russell, & Carroll, 2003) and articulatory lip movements (Green, Nip, Wilson, Mefferd, & Yunusova, 2010). The acoustic properties of this register make it easily discriminable from ADS for adults and infants (Kitamura & Burnham, 2003; Cooper & Aslin, 1990), and infants in the first year of life already show surprise when IDS is used in adult–adult interactions (Soley & Sebastian-Galles, 2020).

No previous research has directly investigated whether adults show increased listening effort or experience a decrease in intelligibility when processing IDS, but there is evidence for differences in cortical responses to this register indicating an increase in attention when adults hear IDS (Peter et al., 2016). IDS presents a unique case. Although it might be speculated that processing IDS would be no different from processing ADS because adults produce IDS naturally when interacting with infants, adults are less accustomed to listening to IDS and so it is also possible that the exaggerated properties of IDS may catch an adult listener off guard and inflict processing costs on the listener. Adults spontaneously adopt IDS and produce it effortlessly in the presence of an infant, but the use of this register in adult–adult interactions would violate interlocutors' social expectations and may even elicit negative reactions (e.g., perceived disrespect). This article is concerned with the relative extent of AV speech benefit in adults' cortical tracking of IDS and ADS.

The AV Speech Benefit

As mentioned above, this study will assess the AV speech benefit as an index of listening effort in situations when listeners are presented with speech styles that are expected

versus unexpected for the communicative context, specifically when IDS and ADS are used to address adults. For this, we will use electrophysiology to measure cortical tracking of speech. Although the bulk of neurophysiological evidence on the AV speech benefit comes from studies that used ERPs (e.g., Baart, Vroomen, Shaw, & Bortfeld, 2014; Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014), more studies are now using cortical tracking, the temporal alignment between neural signals and speech (e.g., Ding & Simon, 2012a, 2012b, 2013; Peelle & Davis, 2012; Ahissar et al., 2001), to examine the AV speech benefit. Unlike ERP studies that entail the use of repetitive and discrete stimuli not representative of natural speech, cortical tracking allows the use of continuous speech stimuli that better represent natural speech. The few research studies conducted to investigate the AV speech benefit in cortical tracking suggest that seeing a speaker's talking face augments cortical tracking of the speech envelope of the speaker (Crosse et al., 2015). For instance, it has been found that tracking of the speaker's temporal speech envelope is stronger when the audio recordings are paired with matching video recordings of the speaker's talking face than when the audio recordings are presented alone (Zion Golumbic, Cogan, Schroeder, & Poeppel, 2013). In addition, when participants are instructed to attend to only one of two speakers presented simultaneously, augmentation of cortical tracking is evident when the speakers are presented in AV than AO mode (Zion Golumbic et al., 2013). Building on these findings, Park, Kayser, Thut, and Gross (2016) found evidence of cortical tracking of the lip movements irrespective of the continuous speech signal and that this phenomenon is stronger when lip movements are congruent to the auditory speech signal compared with when lip movements are incongruent. Furthermore, the researchers also found that stronger cortical tracking of the lip movements during congruent speech supports better speech comprehension. Together, these neurophysiological findings are consistent with behavioral findings indicating that visual speech information benefits speech perception especially in noisy environments (e.g., Sumbly & Pollack, 1954) and that dynamic lip movements facilitate speech processing (Grant & Seitz, 2000). Enhanced tracking of the speaker's speech envelope was also found to be greater in noise than when AV speech is presented in quiet (Crosse, Di Liberto, Bednar, et al., 2016).

The Effects of Gaze Behavior on Speech Perception

Turning to the focus of listeners' attention in the AV speech benefit, when identifying intonation patterns, adults attend more to the upper half of the face, but when identifying words, the same adults shift their focus to the lower half of the face (Lansing & McConkie, 1999). Adults also look more to the eyes than the mouth when presented with questioning expressions, whereas they attend more to the mouth when presented with focused and neutral

expressions (Simonetti, Kim, & Davis, 2016). Thus, the facial regions on which individuals fixate are strongly dependent on the type of information that individuals seek.

Variability in gaze behavior toward the eye and mouth regions influences speech perception. This is clearly illustrated by the between-participants variability in susceptibility to the McGurk effect. Adults who perceived the McGurk effect (a "da" or "tha" response when presented with auditory /ba/ dubbed onto visual /ga/) on at least half the trials had fixation durations predominantly localized on the talker's mouth (Gurler, Doyle, Walker, Magnotti, & Beauchamp, 2015). In contrast, adults who perceived the McGurk effect on less than half the trials fixated more on the talker's eyes. In addition, individuals who looked longer to the mouth consistently reported a McGurk effect and were more likely to make use of visual speech information (Gurler et al., 2015). Thus, individual variability in looking times to the eye and mouth regions can result in differences in AV speech perception.

Interestingly, despite evidence that individual differences in gaze behavior can account for differences in AV speech perception, examinations of gaze behavior to a face have mostly been conducted separately from AV speech perception research. This study bridges this gap by directly investigating the relationship between gaze behavior and AV speech benefit.

This Study

Although IDS is produced naturally by adults when talking to infants, the exaggerated acoustic and visual properties of IDS deviate from an adult listener's expectations and may thus be processed differently from ADS as suggested by previous findings (Peter et al., 2016). This study thus aims to examine whether adults experience an AV speech benefit when watching a speaker address them in IDS compared with ADS and whether this benefit is the same or greater than any AV speech benefit for ADS. This study also aims to investigate whether the AV speech benefit is modulated by gaze behavior.

For these purposes, eye-tracking and EEG data were simultaneously recorded from native Australian English adults as they listened to IDS and ADS segments that were presented in auditory (A), visual (V), and AV modalities. Cortical tracking was indexed by the prediction accuracy of linear models that describe the mapping between the speech envelope and the EEG signal, commonly referred to as temporal response functions (TRFs; e.g., O'Sullivan, Lim, & Lalor, 2019; Crosse, Di Liberto, & Lalor, 2016; Crosse et al., 2015). As in previous studies (e.g., Tan, Kalashnikova, Di Liberto, Crosse, & Burnham, 2022; Crosse, Di Liberto, & Lalor, 2016; Crosse et al., 2015), AV speech benefit was quantified using the additive criterion, that is, [AV vs. (A + V)]; greater cortical tracking in the AV speech condition that deviates from the algebraic sum of that in AO and visual-only (VO) speech [AV > (A + V)] would indicate an AV speech benefit, whereas equivalence would suggest the absence of a multisensory speech benefit.

Although no study has examined whether listening effort is different for IDS and ADS, it is hypothesized that the deviations from ADS will lead to greater effort in processing IDS as in the case of accented speech (e.g., Major et al., 2002; Munro & Derwing, 1995). This possibility is also hinted at by the finding that adults show different cortical responses to auditory IDS and ADS (Peter et al., 2016). As IDS is expected to require greater listening effort than ADS, the AV speech benefit is expected to be greater in IDS than in ADS. Next, as gaze behavior reflects an information-seeking strategy (e.g., Simonetti et al., 2016), participants are expected to attend to the speaker's mouth more during IDS than ADS. Finally, based on previous behavioral evidence that looking behavior modulates AV speech perception (e.g., Gurler et al., 2015), we also expected a positive correlation between participants' attention to the speaker's mouth and the extent of AV speech benefit for both speech types.

METHODS

Participants

A final sample of 16 Australian English monolingual adults aged between 18 and 56 years was included (mean age = 22.76 years, $SD = 9.58$ years, 14 women). This sample was derived from the 18 participants reported in Tan and colleagues (2022) because 16 participants had EEG and eye-tracking data that fulfilled the inclusion criteria (Tan et al., 2022) for both IDS and ADS stimuli. This sample size was modeled on previous work on cortical tracking of the speech signal (e.g., Fiedler, Wöstmann, Herbst, & Obleser, 2019; Hausfeld, Riecke, Valente, & Formisano, 2018; Zion Golumbic et al., 2013). Of the 16 participants, 13 were reported to be right-handed. An additional eight adults were tested but excluded because seven had insufficient gaze data (see Gaze Measures subsection), and one experienced technical failure.

All participants had self-reported normal hearing and normal or corrected-to-normal vision, were free of neurological diseases, and provided written informed consent. The participants took part in this study as part of a Psychology course requirement and received research participation points. This study was approved by the Human Research Ethics Committee at Western Sydney University (Approval Number H11517). The study adhered to the approved protocol regarding participant recruitment, data collection, and data management.

Stimuli

Two sets of 30 short speech passages were recorded, one in IDS (reported in Tan et al., 2022) and one in ADS. The only difference between the content of the IDS and ADS stimuli was the omission of the word "baby" from two of the ADS speech passages (1 and 15; Appendix). IDS and ADS utterances represented the prosodic properties of the two registers: IDS utterances were significantly slower, had higher mean pitch, and had greater pitch range than ADS utterances as it is shown in Table 1.

Sixty (30 passages \times 2 registers) AV recordings were made by a female native speaker of Australian English experienced in producing IDS and ADS. The recordings consisted of a close-up of the speaker's face and shoulders against a white background. There were three presentation modes: AO, VO, and AV. Unimodal auditory and visual recordings were extracted separately from the AV recordings to form AO and VO conditions. In the AO condition, a still image of the speaker's resting face was displayed on the screen as the auditory track was played to control for luminance. In the VO condition, the dynamic video of the speaker's talking face was presented in silence. In the AV condition, both the dynamic video and its sound-track were played together. The auditory recordings were sampled at 44.1 kHz with a 16-bit resolution.

IDS and ADS were presented in two separate blocks, with the order of IDS and ADS presentations counter-balanced across participants. Within each speech type, stimulus presentation order was randomized across modalities, in such a manner that the same sentence did not appear in two modalities on consecutive trials.

Procedure

Participants were informed before the start of the experiment that they were part of a control group for an infant and child study (Tan et al., 2022). They were also told that they would be presented with videos of a speaker talking and were instructed to pay attention to the videos. Participants sat approximately 70 cm away from the center of an LCD screen. Their gaze data were recorded using a Tobii X120 eye tracker positioned below the screen, and their continuous EEG data were recorded with a 128-channel Hydrocel Geodesic Sensor Net, NetAmps 300 amplifier, and NetStation 4.5.7 software (EGI Inc) at a sampling rate of 1 kHz, with the reference electrode placed at Cz. Electrode impedances were kept below 50 k Ω . The EEG recordings were saved for offline analyses. Stimulus

Table 1. Average (SD) Duration, Mean F0, and F0 Range for the Utterances Used as Auditory Stimuli in IDS and ADS

Measure	IDS	ADS	<i>t</i> Test
Duration (msec)	1696.88 (659.37)	1314.93 (587.98)	$t(122) = 20.599, p < .001, d = 1.857$
Mean F0 (Hz)	245.98 (35.80)	221.29 (34.30)	$t(122) = 7.527, p < .001, d = 0.679$
F0 Range (Hz; Max F0–Min F0)	294.01 (100.35)	253.61 (80.53)	$t(122) = 4.026, p < .001, d = 0.363$

presentation was controlled using Presentation software (Neurobehavioural Systems). Triggers indicating the start and end of each trial were recorded along with the EEG. Eye-tracking recordings were co-registered with EEG recordings to ensure that participants were attending to the screen and to examine whether gaze behavior modulates cortical tracking of the speech envelope.

Data Processing

The preprocessing and data analysis pipeline used here was identical to that reported in Tan and colleagues (2022).

Preprocessing

EEG data were preprocessed using EEGLAB (Delorme & Makeig, 2004), FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011), NoiseTools (<https://audition.ens.fr/adc/NoiseTools/>), mTRF Toolbox (Crosse, Di Liberto, Bednar, et al., 2016), and custom scripts in MATLAB R2019a (MathWorks Inc.). EEG data from the three outer rings of the net were removed for our comparison with infants and children (Tan et al., 2022) because these channels have been found to be very noisy in infants and children (Di Liberto et al., 2018; Kalashnikova et al., 2018; Folland, Butler, Payne, & Trainor, 2015). Data from the remaining 92 channels were high-pass filtered at 0.1 Hz and low-pass filtered at 12 Hz with Butterworth 8th order filters. An artifact subspace reconstruction (Kothe & Jung, 2014) was applied to remove noise. Finally, EEG data were rereferenced to the average of all channels (e.g., Kalashnikova et al., 2018) and later downsampled to 100 Hz to reduce processing time.

To investigate the impact of visual speech cues on the cortical tracking of auditory speech, preprocessing of the speech stimuli followed the method described in Jessen, Fiedler, Münte, and Obleser (2019). The auditory soundtracks of each video were extracted, downsampled to 100 Hz to match the sampling rate of the EEG data, and characterized using the broadband speech envelope of the acoustic signal using the NSL toolbox, which models the auditory peripheral and subcortical processing stages (Ru, 2001). The broadband temporal envelope of each soundtrack was obtained by summing up the band-specific envelopes across all 128 logarithmically spaced frequency bands between 0.1 and 4 kHz.

EEG Analysis

Cortical tracking of the speech envelope was measured by estimating the linear mapping between the stimulus speech envelope and the corresponding neural responses to AO, VO, and AV stimuli. Here, the stimulus–response mapping function is modeled in the forward direction using the mTRF Toolbox (see Crosse, Di Liberto, Bednar, et al. [2016] for details), that is, the resulting model describes an optimal linear transformation from the stimulus domain to the neural-signal domain. Such a model is fit by conducting a

lagged ridge regression between the envelope and the EEG data while accounting for probable time delays between speech and the corresponding neural response. The regression weights obtained using this procedure are taken as an estimate of the TRF between the envelope and EEG response at each channel. Significant nonzero weights at a given EEG channel reflect time lags where cortical activity is related to stimulus encoding (Haufe et al., 2014).

Instead of computing individual response functions for each participant, the subject-independent approach was used (see Di Liberto & Lalor, 2017). This involves computing an average response function over $n-1$ participants that is then used to predict the EEG signal of the n th participant via a leave-one-out cross-validation. This subject-dependent modeling approach has been shown to yield better results than the subject-dependent modeling approach when applied to short (5-min) EEG recordings from adults (Jessen et al., 2019). In addition, this approach was used to be consistent with our previous study (Tan et al., 2022) that compared part of these data to infant and child data.

TRFs were initially calculated for each condition at time lags between -200 and 1000 msec before selecting a temporal region of the TRF ($0-600$ msec) because no visible response emerged outside of this range. This window was also chosen because TRFs can be interpreted similarly to ERPs (Crosse et al., 2015), and previous research reported a N1/P2 complex in auditory ERPs associated with AV speech that occurs within $100-200$ msec after the onset of an auditory stimulus (e.g., Pilling, 2009; van Wassenhove, Grant, & Poeppel, 2005; Besle, Fort, Delpuech, & Giard, 2004). As Tan and colleagues (2022) also found TRFs up to 400 msec, the window of $0-600$ msec was chosen. A leave-one-out cross-validation using Tikhonov regularization was conducted to assess how well the unseen EEG data could be predicted based on the TRF. The regularization parameter of the ridge regression was set to $\lambda = 100$ for all participants. This value was chosen to mitigate the potential failure of parameter tuning because of the limited amount of data available (for a discussion, see Crosse et al., 2021). The Pearson correlation coefficient between the predicted and original EEG responses at each electrode was computed to quantify prediction accuracy. Correlation values that are significantly greater than zero indicate that EEG data indeed reflect the encoding of the speech envelope. To investigate AV speech benefit ($A + V$), TRFs were computed and compared with AV TRFs in accordance with the additive model criterion (Stein & Meredith, 1993) as was done in previous studies with similar paradigms (e.g., Crosse, Di Liberto, Bednar, et al., 2016; Crosse et al., 2015). The rationale for the additive model criterion is that multisensory integration can be inferred from the differences between cortical responses to multisensory stimuli and the summation of cortical responses to unisensory stimuli, that is, $[AV - (A + V)]$. The validity of using the additive model to index multisensory integration in electrophysiological studies is well established (Besle et al., 2004). As in Crosse and

colleagues (2015), (A + V) TRFs were derived from the algebraic sum of AO and VO TRFs. The AV speech benefit was quantified as the difference in prediction accuracy for AV TRFs relative to A + V TRFs, that is, $[AV > (A + V)]$.

Gaze Measures

Participants' eye movements were recorded using a Tobii X120 eye tracker, which sampled gaze data at a rate of 120 Hz. Areas of interest (AOIs) of equal dimensions (640×340 pixels) covering the top half and bottom half of the speaker's face demarcated the speaker's eye and mouth regions (Figure 1). Gaze data points for each trial were extracted to calculate the proportion of total looking time (PTL) to these AOIs and to index attention as was done in previous investigations of selective attention to a speaker's eyes and mouth (e.g., Morin-Lessard, Poulin-Dubois, Segalowitz, & Byers-Heinlein, 2019):

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

For analysis purposes, participants were required to have at least 10 out of 30 common trials across the three conditions (AO, VO, and AV) with a minimum of 15% attention for each speech type (IDS and ADS) to be included in the final sample. The mean number of trials per condition included in the analyses is 25.25 for IDS and 25.50 for ADS, and the mean levels of attention across conditions are 78.46 for IDS and 81.12 for ADS.

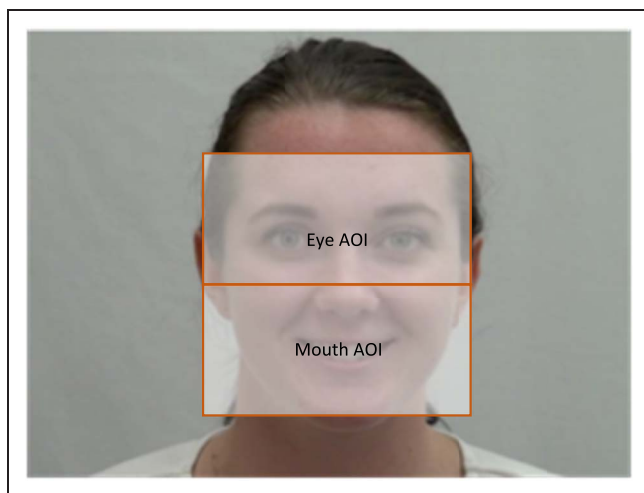


Figure 1. AOIs defined for the speaker's eye and mouth regions.

Statistical Analyses

All statistical analyses were conducted using custom scripts in MATLAB R2020a (MathWorks Inc.).

Prediction accuracies of the TRFs were quantified by Pearson correlation coefficient between the predicted and original EEG responses at each electrode. To examine the presence of cortical tracking across electrodes, mean prediction accuracies across all electrodes were averaged and then tested against zero. To examine whether cortical tracking differed between conditions and speech type, a 2 (Speech Type: IDS vs. ADS) \times 3 (Condition: AO vs. VO vs. AV) repeated-measures ANOVA was conducted with prediction accuracy as the dependent variable. To investigate the AV speech benefit, a 2 (Speech Type: IDS vs. ADS) \times 2 (Condition: AV vs. [A + V]) repeated-measures ANOVA was conducted with prediction accuracy as the dependent variable. To examine gaze behavior, 2 (Speech Type: IDS vs. ADS) \times 3 (Condition: AO vs. VO vs. AV) repeated-measures ANOVAs were conducted with PTL eyes and PTL mouth as dependent variables. In cases where the assumption of sphericity was violated, the Greenhouse–Geisser-corrected degrees of freedom are reported. When required, post hoc comparisons were conducted using two-tailed paired-samples *t* tests with Bonferroni-adjusted alpha levels for multiple comparisons. To examine the relationship between gaze behavior and cortical tracking, Pearson correlation analyses were conducted for each condition between (1) cortical tracking and PTL eyes, and (2) cortical tracking and PTL mouth, where cortical tracking is quantified by TRF prediction accuracy.

RESULTS

Prediction Accuracies

TRF prediction accuracies were tested against zero to assess envelope tracking. Envelope tracking was then compared between conditions and speech types (Figure 2A). Of interest are (1) the differences between cortical tracking of AO, VO, and AV speech in IDS and ADS, (2) the presence of an AV speech benefit as quantified by the additive criterion [i.e., $AV > (A + V)$] in IDS and ADS, and (3) the differences between cortical tracking of IDS and ADS. One-sample *t* tests were first conducted to test prediction accuracies against zero. Next, one-way ANOVAs were conducted for each speech type with their respective prediction accuracies as the dependent variable to examine whether prediction accuracies differed between conditions. Subsequent post hoc comparisons were conducted using two-tailed paired-samples *t* tests with Bonferroni-adjusted alpha levels where multiple comparisons were made.

Evidence of Cortical Tracking of the Speech Envelope

IDS. One-sample *t* tests¹ revealed that prediction accuracy averaged across all electrodes of AO and AV TRFs, AO: $t(15) = 3.14, p = .007$, Hedges' $g = 0.81$; AV: $t(15) = 5.84, p < .001$, Hedges' $g = 1.51$, were significantly greater than

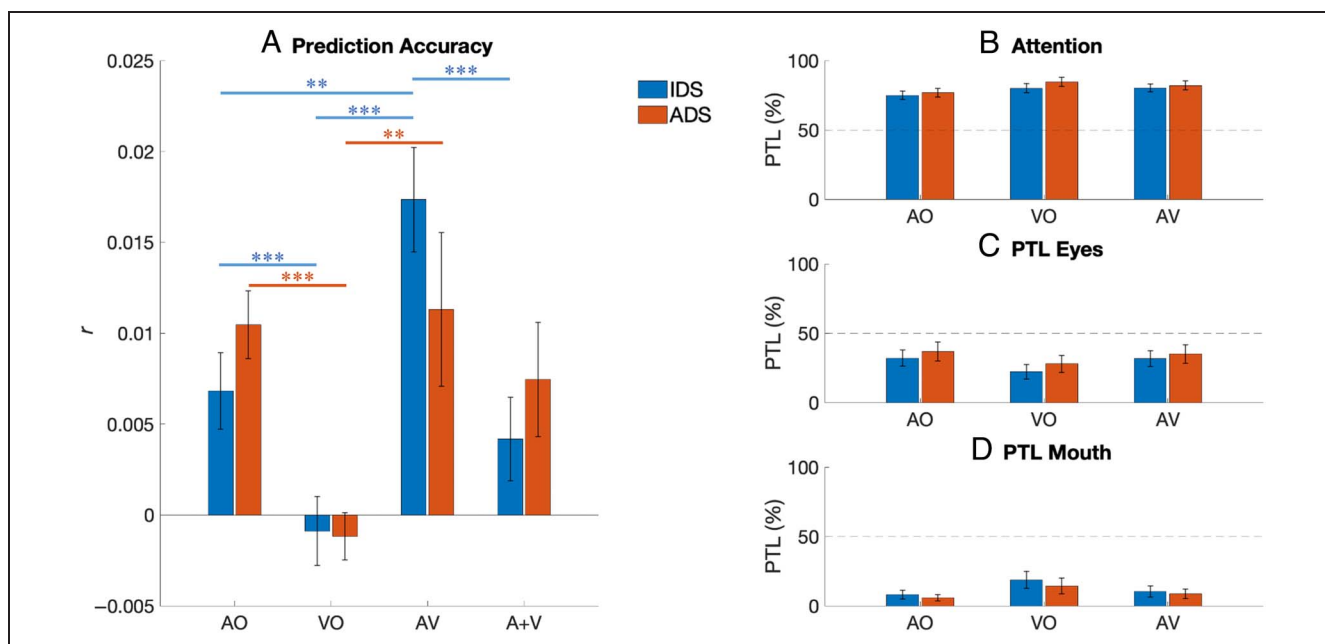


Figure 2. Bar graphs depicting (A) prediction accuracy, (B) attention, (C) PTL eyes, and (D) PTL mouth across conditions and speech types. Error bars represent SEMs. * $p < .05$, ** $p < .01$, *** $p < .001$ for paired-samples t tests with Bonferroni-adjusted alpha levels.

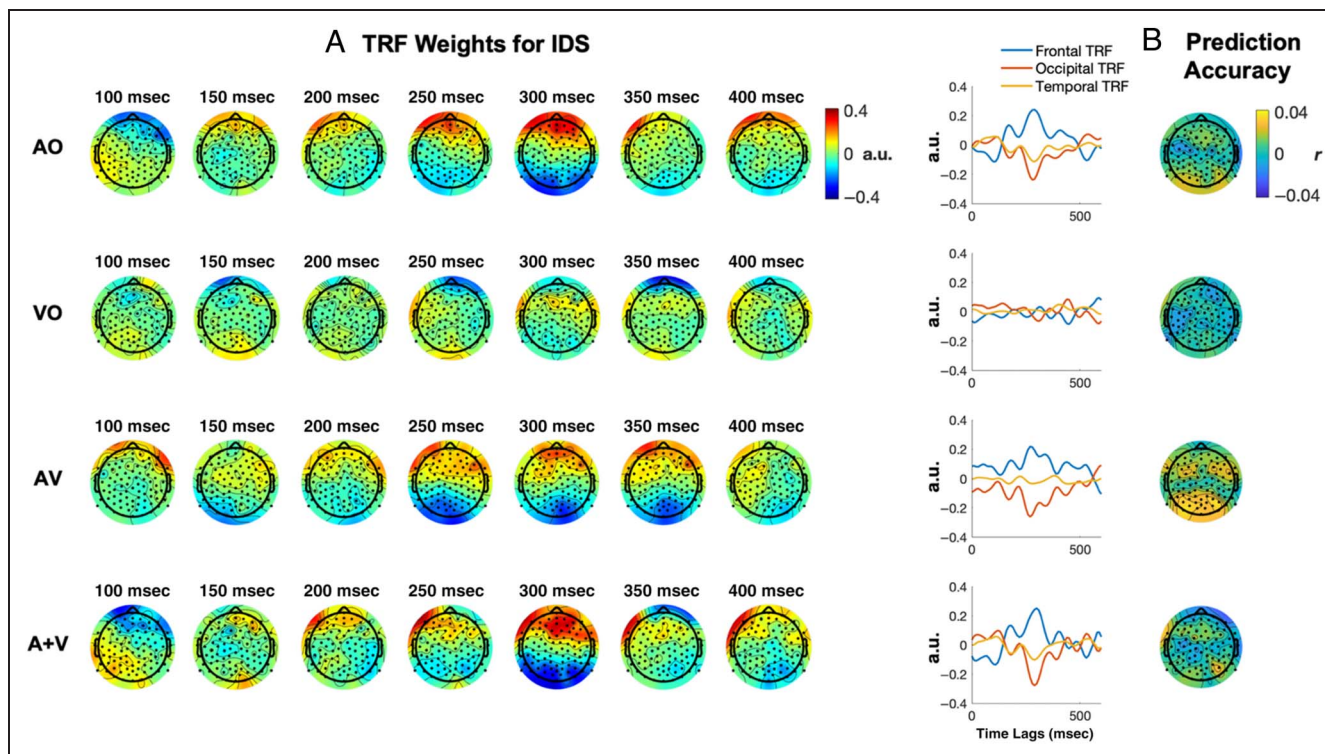


Figure 3. (A) Topographies and TRFs of frontal, occipital, and temporal locations, and (B) prediction accuracy of TRFs for IDS. Topographies are shown from 0 to 400 msec only because TRF peaks occur at around 100/150/200 msec as reported in the speech literature (e.g., Fiedler et al., 2019). Prediction accuracy averaged across all electrodes of AO and AV TRFs was significantly greater than zero (AO: $p = .007$; AV: $p < .001$). When compared between conditions, prediction accuracy of AV TRFs was significantly greater than AO ($p = .001$) and VO TRFs ($p < .001$) and prediction accuracy of AO TRFs was significantly greater than VO TRFs ($p < .001$). Prediction accuracy of AV TRFs was also significantly greater than (A + V) TRFs ($p < .001$), indicating that there was an AV speech benefit for IDS.

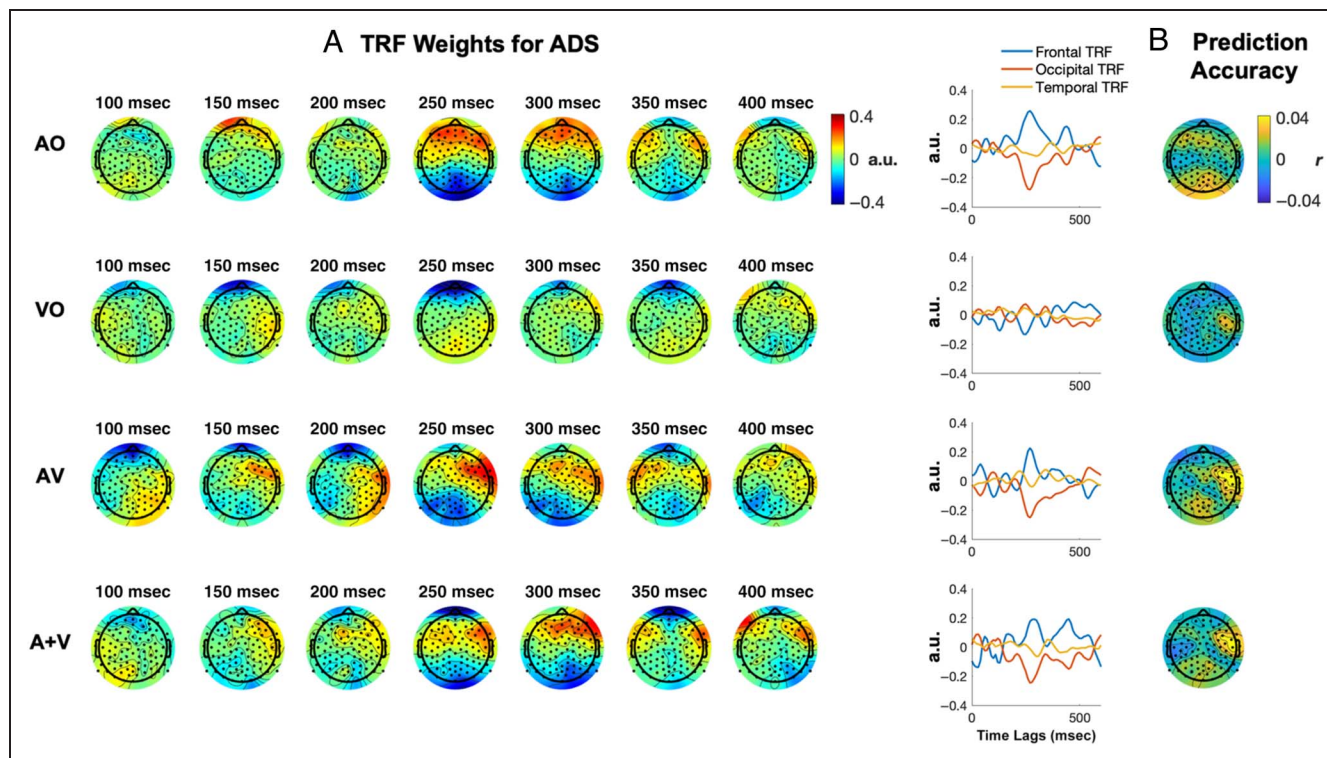


Figure 4. (A) Topographies and TRFs of frontal, occipital, and temporal locations, and (B) prediction accuracy of TRFs for ADS. Prediction accuracy averaged across all electrodes of AO, AV, and (A + V) TRFs was significantly greater than zero (AO: $p < .001$; AV: $p = .02$; A + V: $p = .04$). When compared between conditions, prediction accuracy values of AO and AV TRFs were significantly greater than VO TRFs (AO vs. VO: $p < .001$; AV vs. VO: $p = .008$). Prediction accuracy of AV TRFs was not significantly greater than (A + V) TRFs, indicating that there was no AV speech benefit for ADS.

zero, whereas prediction accuracy of VO TRFs was not, $t(15) = -0.45$, $p = .66$, Hedges' $g = 0.11$. See Figure 3 for topographies of TRF weights and prediction accuracy.

ADS. One-sample t tests indicated that prediction accuracy of AO and AV TRFs were significantly greater than zero, AO: $t(15) = 5.43$, $p < .001$, Hedges' $g = 1.40$; AV: $t(15) = 2.59$, $p = .02$, Hedges' $g = 0.67$, whereas prediction accuracy of VO TRFs was not, $t(15) = -0.88$, $p = .39$, Hedges' $g = 0.23$. See Figure 4 for topographies of TRF weights and prediction accuracy.

IDS versus ADS

A Speech Type (IDS vs. ADS) \times Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted to

examine the differences in cortical tracking of the speech envelopes. The main effect of Speech Type was not significant, $F(1, 15) = 0.16$, $p = .70$, $\eta_p^2 = .01$, whereas the main effect of Condition was significant, $F(2, 30) = 20.73$, $p < .001$, $\eta_p^2 = .58$, and so was the Speech Type \times Condition interaction, $F(2, 30) = 3.61$, $p = .04$, $\eta_p^2 = .14$. The means and standard deviations of prediction accuracy for each speech type and condition are reported in Table 2.

Subsequent post hoc paired-samples t tests with Bonferroni-adjusted alpha level of .016 (.05/3) were conducted to compare the difference in cortical tracking between conditions. For IDS, prediction accuracy of AO TRFs was significantly greater than that of VO TRFs, $t(15) = 4.60$, $p < .001$, Hedges' $g = 0.13$, and prediction accuracy of AV TRFs was significantly greater than that of AO and VO TRFs, AO: $t(15) = 4.05$, $p = .001$, Hedges' $g =$

Table 2. Means (and Standard Deviations) of Prediction Accuracy, Attention, PTL Eyes, and PTL Mouth Across Speech Types and Conditions

	Prediction Accuracy (r)		Attention (%)		PTL Eyes (%)		PTL Mouth (%)	
	IDS	ADS	IDS	ADS	IDS	ADS	IDS	ADS
AO	.007 (.008)	.011 (.008)	74.93 (11.69)	76.73 (12.49)	32.05 (22.96)	36.80 (27.27)	8.05 (12.50)	5.82 (9.39)
VO	-.0009 (.008)	-.001 (.005)	80.11 (13.45)	84.58 (12.82)	22.16 (20.53)	27.87 (24.48)	18.66 (24.38)	14.40 (22.42)
AV	.017 (.012)	.011 (.017)	80.34 (11.14)	82.06 (12.93)	31.54 (22.72)	34.95 (26.19)	10.44 (15.81)	8.75 (13.58)
A + V	.004 (.009)	.007 (.013)	—	—	—	—	—	—

1.05; VO: $t(15) = 7.43, p < .001$, Hedges' $g = 1.87$. For ADS, prediction accuracy of AO TRFs was significantly greater than VO TRFs, $t(15) = 5.36, p < .001$, Hedges' $g = 1.81$, but not significantly different from that of AV TRFs, $t(15) = -0.20, p = .85$, Hedges' $g = 0.06$. Prediction accuracy of AV TRFs was significantly greater than that of VO TRFs, $t(15) = 3.06, p = .008$, Hedges' $g = 1.00$. Comparisons between speech types did not reveal statistically significant differences, AO: $t(15) = 1.33, p = .20$, Hedges' $g = 0.43$; VO: $t(15) = 0.13, p = .90$, Hedges' $g = 0.04$; AV: $t(15) = 1.53, p = .15$, Hedges' $g = 0.39$.

To examine the AV speech benefit, a Speech Type (IDS vs. ADS) \times Condition (AV vs. A + V) repeated-measures ANOVA was conducted with prediction accuracy values as the dependent variable. The main effect of Speech Type was not significant, $F(1, 15) = 0.15, p = .71, \eta_p^2 = .010$, but the main effect of Condition was significant, $F(1, 15) = 12.92, p = .003, \eta_p^2 = .46$. Notably, there was a significant Speech Type \times Condition interaction, $F(1, 15) = 8.52, p = .01, \eta_p^2 = .51$. Subsequent post hoc comparisons between conditions for each speech type revealed that the difference between AV and (A + V) was significant only for IDS, $t(15) = 4.53, p < .001$, Hedges' $g = 1.16$, but not for ADS, $t(15) = 1.38, p = .19$, Hedges' $g = 0.24$. These results indicate that participants experienced an AV speech benefit during IDS but not ADS.

Looking Behavior

Attention

As a preliminary analysis, attention was compared between speech types, because if attention is significantly different between speech types, then attention may be a potential confound for the gaze pattern analyses. The means and standard deviations of attention for each speech type and condition are reported in Table 2. A Speech Type (IDS vs. ADS) \times Condition (AO vs. VO vs. AV) repeated-measures ANOVA revealed only a significant main effect of Condition, $F(0.85, 12.72) = 15.88$ with Greenhouse–Geisser correction, $p < .001, \eta_p^2 = .51$. Subsequent paired-samples t tests revealed that attention during AO trials was significantly lower than VO and AV trials, AO vs. VO: $t(31) = -5.03, p < .001$, Hedges' $g = 0.51$; AO vs. AV: $t(31) = -4.24, p < .001$, Hedges' $g = 0.44$, whereas attention between VO and AV trials was not significantly different, $t(31) = 1.24, p = .23$, Hedges' $g = 0.89$. Neither the main effect of Speech Type nor the Speech Type \times Condition interaction were significant, Speech Type: $F(0.85, 6.36) = 1.16$ with Greenhouse–Geisser correction, $p = .30, \eta_p^2 = .07$; Speech Type \times Condition: $F(0.85, 12.72) = 0.98$ with Greenhouse–Geisser correction, $p = .39, \eta_p^2 = .06$, indicating that participants attended to IDS and ADS trials similarly.

PTL Eyes

A Speech Type (IDS vs. ADS) \times Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted to

examine whether the amount of time spent looking at the speaker's eye region differed between speech types and conditions. The main effect of Condition was significant, $F(1.12, 16.80) = 6.96$ with Greenhouse–Geisser correction, $p < .001, \eta_p^2 = .32$, whereas the main effect of Speech Type, $F(0.56, 8.40) = 2.18$ with Greenhouse–Geisser correction, $p = .16, \eta_p^2 = .13$, and the Speech Type \times Condition interaction were not significant, $F(1.12, 16.80) = 0.48$ with Greenhouse–Geisser correction, $p = .62, \eta_p^2 = .031$. Paired-samples t tests were conducted to examine the main effect of Condition. PTL eyes was greater in AO and AV conditions compared with the VO condition, AO vs. VO: $t(31) = 3.97, p < .001$, Hedges' $g = 0.38$; AV vs. VO: $t(31) = 4.35, p < .001$, Hedges' $g = 0.33$, but PTL eyes was not significantly different during AO and AV trials, $t(31) = 0.60, p = .55$, Hedges' $g = 0.05$ (Figure 2B). The means and standard deviations of PTL eyes for each speech type and condition are reported in Table 2.

PTL Mouth

A Speech Type (IDS vs. ADS) \times Condition (AO vs. VO vs. AV) repeated-measures ANOVA was conducted to examine whether PTL mouth differed between speech types and conditions. The main effect of Speech Type and the main effect of Condition were significant, Speech Type: $F(0.32, 4.86) = 5.09$ with Greenhouse–Geisser correction, $p = .039, \eta_p^2 = .25$; Condition: $F(0.65, 9.72) = 6.38$ with Greenhouse–Geisser correction, $p = .015, \eta_p^2 = .30$, but the Speech Type \times Condition interaction was not significant, $F(0.65, 9.72) = 1.95$ with Greenhouse–Geisser correction, $p = .18, \eta_p^2 = .12$. Paired-samples t tests indicated that PTL mouth was significantly greater during VO compared with AO and AV trials, VO vs. AO: $t(31) = 3.69, p < .001$, Hedges' $g = 0.51$; VO vs. AV: $t(31) = 3.78, p < .001$, Hedges' $g = 0.35$, whereas PTL mouth during AO and AV trials was not significantly different, $t(31) = -2.00, p = .054$, Hedges' $g = 0.20$ (Figure 2C). The main effect of Speech Type indicates that PTL mouth was greater during IDS compared with ADS trials. As the Speech Type \times Condition interaction was not significant, no further post hoc tests were conducted. The means and standard deviations of PTL mouth for each speech type and condition are reported in Table 2.

Relationship between Prediction Accuracy and Gaze Measures

Pearson correlations were conducted to investigate whether gaze measures (PTL eyes and PTL mouth) were related to prediction accuracy. None of the correlations were significant (IDS: all $r_s < .13$, all $p_s > .48$; ADS: all $r_s < .22$, all $p_s > .42$).

Additional exploratory correlational analyses between attention and prediction accuracy were conducted to clarify these unexpected nonsignificant results. These

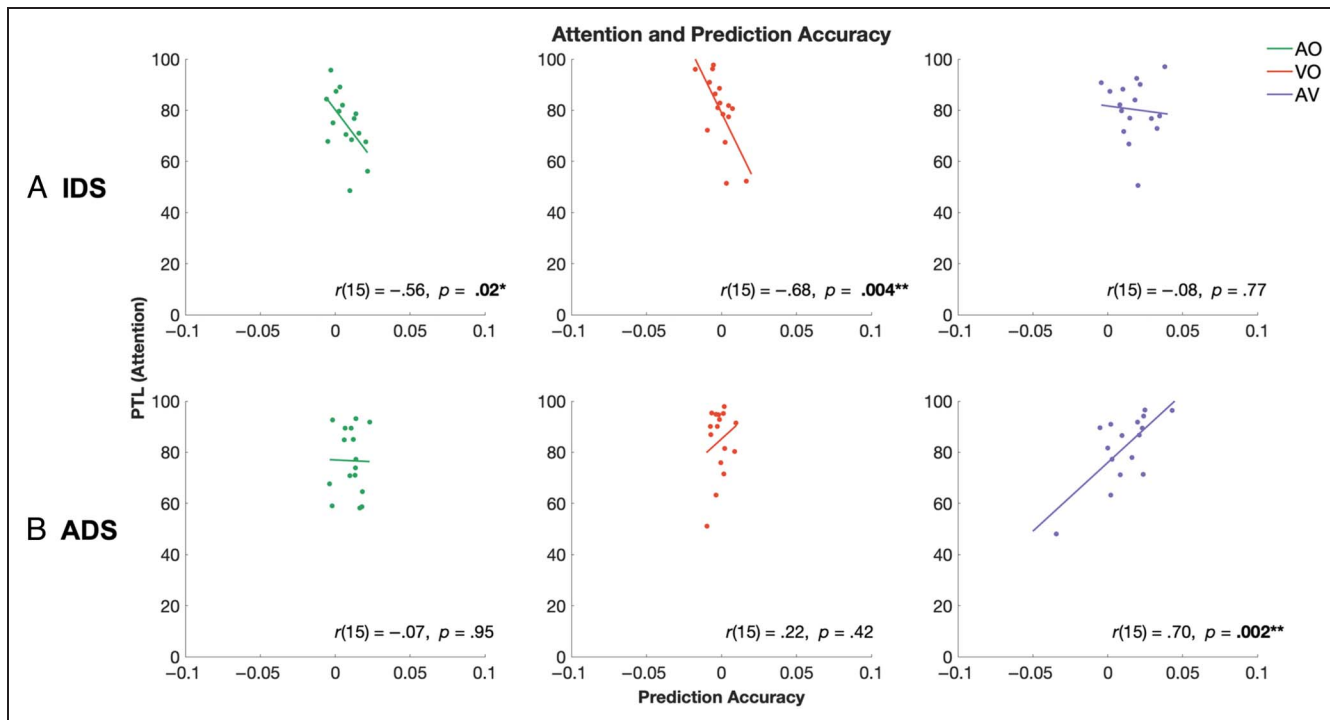


Figure 5. Scatter plots and correlations between attention and prediction accuracy for (A) IDS and (B) ADS.

analyses revealed that prediction accuracy of AO and VO TRFs in IDS were negatively correlated with attention, AO: $r(15) = -.55, p = .02$; VO: $r(15) = -.68, p = .004$, whereas the prediction accuracy of AV TRFs in ADS was positively correlated with attention, $r(15) = .70, p = .002$ (Figure 5).

DISCUSSION

To examine the relationship between looking behavior and cortical tracking of speech, eye-tracking and EEG data were simultaneously recorded as adults were presented continuous IDS and ADS segments in auditory, visual, and AV modalities. Analyses indicated that cortical tracking of the speech envelope was significant in AO and AV modalities for both IDS and ADS. Notably, participants showed an AV speech benefit during IDS but not during ADS trials. In IDS, cortical tracking accuracy of AV trials was greater than AO trials, whereas for ADS, cortical tracking accuracy of AV trials was not different from AO trials. In addition, gaze behavior to the eyes and mouth differed between IDS and ADS. Finally, gaze measures (i.e., PTL eyes and PTL mouth) were not correlated with cortical tracking of IDS and ADS. These three issues—the AV speech benefit, gaze behavior, and cortical tracking-gaze behavior correlations are considered in turn below.

An AV Speech Benefit for IDS But Not ADS

In line with previous research (e.g., Crosse, Di Liberto, Bednar, et al., 2016; Ding & Simon, 2014; Giraud &

Poepfel, 2012), adults in this study showed cortical tracking of the speech envelope in both AO and AV conditions. Most importantly, there was an AV speech benefit for IDS but not for ADS. Although the exaggerated properties of IDS enhance speech processing for infants (e.g., Kalashnikova et al., 2018), it is possible that the same properties may hinder speech processing for adults because adults do not expect to be spoken to in IDS, although they can produce it quite naturally.

Although the IDS and ADS stimuli in this study were spoken by the same speaker and were nearly identical in content (except for the use of the word “baby” in IDS but not ADS), they differed significantly in acoustic properties. Thus, although the passages produced in the two registers should have been equally easy to comprehend, our findings suggest that the properties of the auditory ADS signal alone were sufficient for adults’ speech processing without the aid of visual speech information, but in IDS, an additional factor over and above the speech content presumably made the IDS passages less easy to process.

In difficult listening conditions, visual speech cues especially enhance speech perception. Behavioral studies report a near-ubiquitous AV speech benefit in quiet (e.g., Fort et al., 2013; Navarra & Soto-Faraco, 2007) and in noise (e.g., Moradi et al., 2013; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Schwartz et al., 2004; Erber, 1969) and that the extent of this benefit is also larger during adverse listening conditions (Ross et al., 2007). In this regard, Crosse, Di Liberto, Bednar, and colleagues (2016) provided neurophysiological evidence for greater

enhancement of tracking of the speaker's speech envelope when AV speech was embedded in noise than when AV speech was presented in quiet. As in Crosse, Di Liberto, Bednar, and colleagues (2016), adults in the current study showed an AV gain during an atypical listening condition (IDS). Therefore, our findings suggest that visual speech cues may ameliorate the challenges that unimodal IDS imposes on adults' speech perception because such cues are temporally synchronous with the auditory signal (Bahrick & Lickliter, 2002) or may be complementary in that they help disambiguate highly confusable phonemes (Potamianos, Neti, Gravier, Garg, & Senior, 2003). Although the evidence is consistent with IDS entailing difficult listening conditions for adult listeners, further work that includes comprehension tests and questionnaires on adults' perception of IDS stimuli is required.

One hypothesis for the neural mechanisms underlying the AV speech benefit is that the onset of visual speech cues resets the phase of ongoing oscillations in the auditory cortex (Mercier et al., 2015; Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008). Articulatory information from lip movements correlates strongly with their corresponding speech sounds (Chandrasekaran et al., 2009, but see Schwartz & Savariaux, 2014), and the degree of syllable visual predictability influences functional connectivity between the auditory and visual cortices (Arnal, Morillon, Kell, & Giraud, 2009). So, it is postulated that the phase-reset puts the auditory cortex in a receptive and excitable state (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007) that allows for predictions of the upcoming auditory signal to be encoded (Arnal, Wyart, & Giraud, 2011) and for the predicted input to be processed more easily (Henry & Obleser, 2012; Friston, 2010). Support for this hypothesis comes from findings that visual speech information phase-resets the low-frequency neuronal oscillations in the auditory cortex and that this phase-reset amplifies the strength of cortical responses to the auditory signal (Mégevand et al., 2020). In addition, the amount of predictive information provided by visual speech cues is positively correlated with the degree of AV speech benefit experienced (van Wassenhove et al., 2005). As an example, visual speech cues that are related to the place of articulation will provide predictive information because these cues can be observed readily from the speaker's articulatory movements and are not affected by background noise (Grant & Bernstein, 2019). The presence of an AV speech benefit only for IDS and not for ADS in our study suggests that visual speech cues played a larger role in processing IDS than in ADS, hinting at the possibilities that the degree of phase-reset is lower for IDS compared with ADS, and/or that the extent of phase alignment before phase-reset is lower for IDS than for ADS. Nevertheless, further research is required because the exact neuronal mechanisms underlying the AV speech benefit remain unclear although there are extensive behavioral and neurophysiological

findings that visual speech information modulates speech perception.

The AV Speech Benefit and Gaze Behavior

The AV speech benefit is assumed to rely on individuals' attention to a speaker's talking face to pick up on available visual speech information. Previous studies have found that adults generally fixate on the speaker's eyes (Lewkowicz & Hansen-Tift, 2012) but shift their focus to the speaker's mouth when listening conditions become difficult or when the auditory signal is unreliable (Birulés, Bosch, Pons, & Lewkowicz, 2020; Yi, Wong, & Eizenman, 2013). The same pattern is found in this study: Participants attended to the mouth more during IDS than ADS and during VO trials than AO and AV trials in both IDS and ADS. This pattern of gaze behavior is argued to be part of an information-seeking strategy and is in accordance with the cognitive relevance hypothesis (Henderson, Malcolm, & Schandl, 2009), which postulates that visual attention is driven by current information-gathering needs more than by visual salience. For instance, Lusk and Mitchel (2016) found that, when presented with AV recordings of a speaker producing trisyllabic words from an artificial language, adults attended most to the speaker's mouth, but their attention to the mouth decreased as they became more familiar with the artificial language.

If, according to the cognitive relevance hypothesis, attention to the mouth is an information-seeking strategy, then one would expect a positive correlation between cortical tracking accuracy and PTL mouth. However, correlational analyses conducted to examine the relationship here between cortical tracking accuracy and looking behavior to the eye and mouth regions were not significant. This conflicts with the premise of the cognitive relevance hypothesis that gaze behavior to the speaker's mouth is an information-seeking strategy. Other studies have also failed to find a relationship between adults' looking behavior and speech recognition performance (Buchan, Paré, & Munhall, 2008; Buchan et al., 2007), so the present results are not entirely unexpected. At first glance, this calls into question the utility of such an information-seeking strategy in which adults redirect their attention to the speaker's mouth region especially in difficult listening conditions. However, it is also possible that the looking behavior does not accurately index the AV speech benefit. Recent findings that the extent to which listeners benefit from the presence of visual speech compared with increased acoustic clarity is correlated with their lip-reading ability (Aller, Økland, MacGregor, Blank, & Davis, 2022) suggests that a listener's lip-reading ability—rather than the looking behavior to the speaker's mouth per se—may reflect more precisely their capacity to derive an AV speech benefit.

Exploratory correlational analyses were conducted to clarify the null results from correlational analyses between

gaze measures (PTL eyes and PTL mouth) and cortical tracking accuracy. These analyses revealed a significant negative correlation between the overall attention to the visual display and cortical tracking accuracy of AO and VO trials in IDS, and a significant positive correlation between attention and cortical tracking accuracy of AV trials in ADS. These correlations provide additional insights when considered alongside the correlational analyses conducted for PTL eyes and PTL mouth.

The positive correlation between attention and cortical tracking accuracy of AV trials in ADS provides indirect support for previous AV speech benefit research by showing that gaze behavior is associated with cortical tracking of continuous AV speech but suggests that general gaze behavior to the screen, the whole display—and not gaze behavior to a particular region of the speaker’s face—is sufficient for acquiring linguistic information during speech processing. Indeed, visual speech information gleaned from the periphery can still be accurately processed (Paré, Richler, ten Hove, & Munhall, 2003), suggesting that direct eye-gaze patterns may not necessarily entirely index all the visual information that has been processed because they do not include information processed in peripheral vision.

The negative correlation between attention and cortical tracking found in VO in IDS suggests that visual speech cues from IDS may hinder rather than aid speech perception, and this is quite possibly because IDS is also conveyed through exaggerated facial expressions (Chong et al., 2003), which are unexpected in speech directed to adult listeners. Drawing on previous findings that eyebrow and rigid head movements are related to changes in fundamental frequency (Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Cavé et al., 1996) and prosody (Cvejic, Kim, & Davis, 2010; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004), and that inaccurate articulatory information by nonnative English speakers hampers phoneme recognition by native listeners (Kawase et al., 2014), it could be speculated that the exaggerated head and lip movements in the silent videos in IDS VO trials (devoid of their auditory concomitants) were detrimental to participants’ speech perception. The presumed oddity of visual speech cues involved in IDS to adults (especially when highlighted in VO speech) may compound the costs involved in processing the auditory signal of IDS. Speech perception is inherently predictive: Listeners process the incoming speech signal while actively anticipating upcoming information (Clark, 2013) to facilitate and reduce the processing load (Pickering & Garrod, 2007). These predictive mechanisms are affected when acoustic properties deviate from expectations—as in foreign-accented speech and presumably in IDS (Räsänen, Kakouros, & Soderstrom, 2018). Listeners of foreign-accented speech entertain more candidate words (i.e., a larger lexical-semantic neighborhood) for longer durations even after successful comprehension (Porretta & Kyröläinen, 2019), likely because of the uncertainty of the auditory signal (Brouwer

& Bradlow, 2016). Further evidence for the adverse effects of atypical speech on speech processing comes from reduced effectiveness of auditory primes as a speaker’s accent becomes thicker (Porretta, Tucker, & Järvikivi, 2016), suggesting that the predictive mechanisms are implicated when the speech signal differs from what is typically experienced (Porretta et al., 2020).

A possible explanation for the negative correlation between prediction accuracy and attention in the AO condition in IDS is that participants were inattentive although they were looking at the screen. It may be that the auditory recordings and the static photo of the speaker’s face are not as striking as the dynamic AV or VO conditions. Although participants were instructed to listen attentively to the speaker, there may have been instances where their minds wandered during AO trials. This is supported by research showing that mind wandering in adults tends to produce long fixations (e.g., Krasich et al., 2018; Bixler & D’Mello, 2016; Foulsham, Farley, & Kingstone, 2013) and is associated with increased looking durations on the speaker’s image (Zhang, Miller, Sun, & Cortina, 2020). To clarify whether this is indeed the case, future studies could include a self-report measure of mind wandering and, more objectively, a comprehension test at the end as a measure of attention. Future work could also compare passages differing in comprehension difficulty to modulate inattention.

Limitations and Future Directions

The explanations raised here regarding the correlational results between cortical tracking and looking behavior remain speculative and await future research. To determine whether participants perceive visual speech cues of IDS as unexpected, odd, and/or distracting, future studies could include a short questionnaire, and participants could also be asked to report their familiarity with IDS. For instance, participants who have infants in their households may be more familiar with IDS and may show stronger cortical tracking accuracy of IDS than participants who have almost no exposure to or experience in producing IDS. Next, to determine whether visual speech cues of IDS impede the predictive mechanisms underlying speech perception, future studies could examine IDS by implementing a gating paradigm as used in previous AV speech perception studies (e.g., Moradi, Lidestam, Saremi, & Rönnerberg, 2014; Moradi et al., 2013). In these, participants could be presented AV sentences in increasing gates until a target is correctly identified. If IDS is more challenging to process than ADS, then participants should require more time for correct identification of IDS targets than ADS targets. Such a study would directly address the issue of whether there are costs involved in adults’ processing of IDS.

Although the present study did not explicitly measure listening effort, future studies should also include a rating scale for participants to rate their perception of the degree

of listening effort required for IDS and ADS. This measurement would then allow for direct conclusions to be drawn between participants' perceived listening effort and their behavioral performance in AV speech perception tasks. Including an additional physiological measure such as pupil dilation (de Gee, Correa, Weaver, Donner, & van Gaal, 2021) or skin conductance response (Jang, Park, Park, Kim, & Sohn, 2015) as a marker of surprise may aid in teasing apart whether IDS deviates from expectations and whether this deviation modulates listening effort.

Another possible future direction is to compare the extent of AV speech benefit experienced in different listening conditions. Most research investigating speech perception in suboptimal listening conditions uses paradigms that involve vocoded speech (Bernstein, Auer, Eberhardt, & Jiang, 2013), multiple speakers (Schwartz et al., 2004; Rudmann et al., 2003), or varying degrees of speech in noise (speech-to-noise ratio; Moradi et al., 2013). Varying acoustic speech intelligibility in a more naturalistic manner (as opposed to vocoded speech) and comparing speech perception in different suboptimal listening conditions (e.g., IDS, foreign-accented speech, multitalker speech) would allow the context-dependent contribution of visual speech cues to speech perception to be discerned.

Conclusion

As expected, based on the evidence for greater AV speech benefit in cortical tracking when speech is difficult or violates social expectations of the communicative interaction, the AV speech benefit here was present for IDS but not for ADS. This evidence is consistent with the notion that IDS is an unexpected register for adult listeners, but further research is required to pinpoint the source of the AV speech benefit for IDS. In addition, participants attended more to the speaker's mouth during IDS than ADS trials, suggesting that IDS was more challenging to comprehend than ADS. Accordingly, it was expected that there would be a correlation between cortical tracking accuracy and looking behavior to the speaker's mouth; however, there was no evidence of this. Nevertheless, exploratory analyses revealed unexpected negative correlations between attention and cortical tracking of AO and VO trials in IDS. These unexpected and seemingly contradictory findings are preliminary and call for future studies to examine further whether there are processing costs involved in comprehending IDS and, if so, whether such processing costs occur in unimodal speech but are mitigated by multimodal presentations of IDS.

APPENDIX

Stimuli

-
1. Hi baby! How are you today? It's a wonderful day today! You look happy! Are you ready for some fun?
 2. We're going to read a story now! Do you want to read a story? This story is about a sheep, shoe and shark. Here we go!
 3. Here comes the sheep! It's a lovely sheep! It's nice and fluffy! The sheep has white wool! Can you feel the soft wool?
 4. Does the sheep have a tail? Does the sheep have ears? Can you see the two ears? The nice soft sheep has two ears!
 5. This sheep has a black nose! What noise does the sheep make? Ba-a ba-a. You like the sheep, don't you?
 6. Let's have a look at some pictures! What have we got here? This is a funny picture! Should we look at another picture? Look at this one!
 7. What is this? This is a very nice shoe! What colour is the shoe? It's a red shoe. I like red shoes!
 8. Shoes go on feet! Can we put the shoes on your feet? Do you like the shoes? Do you want to hold the shoes?
 9. These are some beautiful flowers! That's a daisy! There's a butterfly on the daisy! Can you see the butterfly?
 10. Can you see the sky? Look at the blue sky! Isn't it beautiful? The clouds look like fairy floss!
 11. Uh oh! Someone threw the shoe up into the tree! Isn't that silly? What should we do now?
 12. Can you look here? Here is the shark! Can you see the shark? The shark is having a good time!
 13. Would you like to clap for me? Clap clap clap! Clap your hands! You love that, don't you?
 14. It's a beautiful day today! I see the sun shining brightly! It's really nice that you came to see me today!
 15. You came from far away to see me today! It's really nice to meet you! I hope it's also the case for you! You are a beautiful baby!
 16. We're going to the park later! The weather is great for some outdoor time! I bet we're going to see many dogs! What do you think?
 17. My! Look at those shoes! There are pink flowers on them! Aren't they pretty? I think I can wear them all day! Do you like them too?
 18. Look! I have four soft toys. How many have you got? My favourite is this brown teddy bear! I take it everywhere with me!
 19. What would you like to do today? It's a very sunny day! Shall we go to the beach? Let's head to the beach! It's time to find our cossies!
 20. There's a little bird over there! It's green and red and yellow! Oh, look! It's eating a worm! The little bird must be hungry!
 21. A rainbow has seven colours! Red, orange, yellow, green, blue, violet, and indigo. My favourite colour is blue! What's your favourite colour?
 22. Can you hear that sound? I wonder where it's coming from! Oh, it's coming from under that box! Will we take a look?
 23. Whee! This is fun! I love playgrounds! My favourite part is the slide! The higher the slide, the better it is! Do you like slides too?
 24. Look! A kitten's coming my way! It has grey patches all over it! It's so cute! I want to play with it! Let's go!
 25. It's me again! Look at what I've got! What do we have here? What's in here? Oh, it's an apple! I love eating apples!
 26. Wow! Look at that tree! It has lots of pretty pink flowers on it! Let's go closer so we can have a better look!
 27. I like going to parks! There's such a huge space to run about! What about you? Do you like parks too?
 28. When the weather gets too hot, I like to wear my cap! The cap keeps the sun off my face. Do you also wear a cap?
 29. When the weather gets too cold, I always drink hot chocolate! It keeps me warm! What's your favourite drink?
 30. You've been so attentive today! Thank you for listening to me! I enjoyed talking to you! I hope you enjoyed listening to me too!
-

Reprint requests should be sent to Jessica Tan, Science of Learning in Education Centre, Office of Education Research, National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore 637616, Singapore, or via e-mail: jessica.tan@nie.edu.sg.

Data Availability Statement

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

Author Contributions

Sok Hui Jessica Tan: Conceptualization; Formal analysis; Investigation; Methodology; Writing—Original draft; Writing—Review & editing. Marina Kalashnikova: Conceptualization; Writing—Original draft; Writing—Review & editing. Giovanni M. Di Liberto: Formal analysis; Methodology; Writing—Review & editing. Michael J. Crosse: Formal analysis; Methodology; Writing—Review & editing. Denis Burnham: Conceptualization; Supervision; Writing—Original draft; Writing—Review & editing.

Funding Information

This research was funded by a doctoral scholarship to J. T. funded by the MARCS Institute at Western Sydney University and the HEARING Cooperative Research Centre, and by HEARING Cooperative Research Centre funding to D. B. M. K. is supported by the Basque Government through the BERC 2018–2021 program and PIBA PI-2019-0054, and by the Spanish Ministry of Science and Innovation (<https://dx.doi.org/10.13039/501100004837>) through the Ramon y Cajal Research Fellowship, grant number: RYC2018-024284-I. G. D. L. is supported by Science Foundation Ireland (<https://dx.doi.org/10.13039/501100001602>) under grant number: 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience (JoCN)* during this period were $M(\text{an})/M = .407$, $W(\text{oman})/M = .32$, $M/W = .115$, and $W/W = .159$, the comparable proportions for the articles that these authorship teams cited were $M/M = .549$, $W/M = .257$, $M/W = .109$, and $W/W = .085$ (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

Note

1. It is possible that values different from 0 may still occur by chance, which would not be captured by one-sample *t* tests against 0 (we thank an anonymous reviewer for pointing this out). To ensure that this was not the case, a permutation test with 100 randomizations was additionally conducted against a null distribution of shuffled data for each condition in both speech types. Shuffled data were derived by randomly shuffling the order of trials in the stimulus. The permutation tests showed the same pattern of results as the one-sample *t* tests: prediction accuracy values of AO (IDS: $p = .015$; ADS: $p = .001$), AV (IDS: $p = .015$; ADS: $p = .006$), and (A + V) (IDS: $p = .011$; ADS: $p < .001$), but not VO (IDS and ADS: $p_s > .05$), were significantly greater than the shuffled data for both IDS and ADS.

REFERENCES

- Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 520–529. <https://doi.org/10.1037/a0013552>, PubMed: 19331505
- Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, 98, 13367–13372. <https://doi.org/10.1073/pnas.201400998>, PubMed: 11698688
- Aller, M., Økland, H. S., MacGregor, L. J., Blank, H., & Davis, M. H. (2022). Differential auditory and visual phase-locking are observed during audio-visual benefit and silent lip-reading for speech perception. *Journal of Neuroscience*, 42, 6108–6120. <https://doi.org/10.1523/JNEUROSCI.2476-21.2022>, PubMed: 35760528
- Alsius, A., Wayne, R. V., Paré, M., & Munhall, K. G. (2016). High visual resolution matters in audiovisual speech perception, but only for some. *Attention, Perception, & Psychophysics*, 78, 1472–1487. <https://doi.org/10.3758/S13414-016-1109-4>, PubMed: 27150616
- Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, 29, 13445–13453. <https://doi.org/10.1523/JNEUROSCI.3194-09.2009>, PubMed: 19864557
- Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14, 797–801. <https://doi.org/10.1038/nn.2810>, PubMed: 21552273
- Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, 92, 339–355. <https://doi.org/10.1348/000712601162220>, PubMed: 11417785
- Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, 130, 31–43. <https://doi.org/10.1016/j.cognition.2013.09.006>, PubMed: 24141035
- Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 30, pp. 153–187). Academic Press. <https://psycnet.apa.org/record/2003-02423-004>, [https://doi.org/10.1016/S0065-2407\(02\)80041-6](https://doi.org/10.1016/S0065-2407(02)80041-6), PubMed: 12402674
- Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Audiovisual cues benefit recognition of accented speech in

- noise but not perceptual adaptation. *Frontiers in Human Neuroscience*, 9, 422. <https://doi.org/10.3389/fnhum.2015.00422>, PubMed: 26283946
- Bernstein, L. E., Auer, E. T., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience*, 7, 34. <https://doi.org/10.3389/fnins.2013.00034>, PubMed: 23515520
- Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20, 2225–2234. <https://doi.org/10.1111/j.1460-9568.2004.03670.x>, PubMed: 15450102
- Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition, & Neuroscience*, 35, 1–12. <https://doi.org/10.1080/23273798.2020.1762905>
- Bixler, R., & D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26, 33–68. <https://doi.org/10.1007/S11257-015-9167-1>
- Bobb, S. C., Mello, K., Turco, E., Lemes, L., Fernandez, E., & Rothermich, K. (2019). Second language learners' listener impressions of foreigner-directed speech. *Journal of Speech, Language, and Hearing Research*, 62, 3135–3148. https://doi.org/10.1044/2019_JSLHR-S-18-0392, PubMed: 31412215
- Bosseler, A. N., Teinonen, T., Tervaniemi, M., & Huotilainen, M. (2016). Infant directed speech enhances statistical learning in newborn infants: An ERP study. *PLoS One*, 11, e0162177. <https://doi.org/10.1371/journal.pone.0162177>, PubMed: 27617967
- Brouwer, S., & Bradlow, A. R. (2016). The temporal dynamics of spoken word recognition in adverse listening conditions. *Journal of Psycholinguistic Research*, 45, 1151–1160. <https://doi.org/10.1007/S10936-015-9396-9>, PubMed: 26420754
- Buchan, J., Paré, M., & Munhall, K. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, 2, 1–13. <https://doi.org/10.1080/17470910601043644>, PubMed: 18633803
- Buchan, J., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behaviour during audiovisual speech perception. *Brain Research*, 1242, 162–171. <https://doi.org/10.1016/j.brainres.2008.06.083>, PubMed: 18621032
- Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, 296, 1435. <https://doi.org/10.1126/science.1069587>, PubMed: 12029126
- Burnham, D., Vatikiotis-Bateson, E., Barbosa, A. V., Menezes, J. V., Yehia, H. C., Morris, R. H., et al. (2022). Seeing lexical tone: Head and face motion in production and perception of Cantonese lexical tones. *Speech Communication*, 141, 40–55. <https://doi.org/10.1016/j.specom.2022.03.011>
- Cavé, C., Guitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and Fo variations. *Proceeding of the Fourth International Conference on Spoken Language Processing. ICSLP '96*, 4 (pp. 2175–2178). <https://doi.org/10.1109/ICSLP.1996.607235>
- Chandrasekaran, C., Trubanova, A., Stillitano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5, e1000436. <https://doi.org/10.1371/journal.pcbi.1000436>, PubMed: 19609344
- Chong, S. C. F., Werker, J. F., Russell, J. A., & Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development*, 12, 211–232. <https://doi.org/10.1002/ICD.286>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences*, 36, 181–204. <https://doi.org/10.1017/S0140525X12000477>, PubMed: 23663408
- Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, 61, 1584–1595. <https://doi.org/10.1111/J.1467-8624.1990.tb02885.x>, PubMed: 2245748
- 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, 14195–14204. <https://doi.org/10.1523/JNEUROSCI.1829-15.2015>, PubMed: 26490860
- Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10, 604. <https://doi.org/10.3389/fnhum.2016.00604>, PubMed: 27965557
- Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, 36, 9888–9895. <https://doi.org/10.1523/JNEUROSCI.1396-16.2016>, PubMed: 27656026
- Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in Neuroscience*, 15, 705621. <https://doi.org/10.3389/fnins.2021.705621>, PubMed: 34880719
- Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, 52, 555–564. <https://doi.org/10.1016/j.specom.2010.02.006>
- de Gee, J. W., Correa, C. M., Weaver, M., Donner, T. H., & van Gaal, S. (2021). Pupil dilation and the slow wave ERP reflect surprise about choice outcome resulting from intrinsic variability in decision confidence. *Cerebral Cortex*, 31, 3565–3578. <https://doi.org/10.1093/cercor/bhab032>, PubMed: 33822917
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134, 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>, PubMed: 15102499
- Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hearing Research*, 348, 70–77. <https://doi.org/10.1016/j.heares.2017.02.015>, PubMed: 28246030
- Di Liberto, G. M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., & Lalor, E. C. (2018). Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. *NeuroImage*, 175, 70–79. <https://doi.org/10.1016/j.neuroimage.2018.03.072>, PubMed: 29609008
- Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107, 78–89. <https://doi.org/10.1152/jn.00297.2011>, PubMed: 21975452
- Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, U.S.A.*, 109, 11854–11859. <https://doi.org/10.1073/pnas.1205381109>, PubMed: 22753470
- Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of

- speech. *Journal of Neuroscience*, *33*, 5728–5735. <https://doi.org/10.1523/JNEUROSCI.5297-12.2013>, PubMed: 23536086
- Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311. <https://doi.org/10.3389/fnhum.2014.00311>, PubMed: 24904354
- Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*, 423–425. <https://doi.org/10.1044/jshr.1202.423>, PubMed: 5808871
- Erdener, D., & Burnham, D. (2013). The relationship between auditory–visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology*, *116*, 120–138. <https://doi.org/10.1016/j.jecp.2013.03.003>, PubMed: 23773915
- Ferguson, S. H., Keum, K. A., Jongman, A., & Sereno, J. A. (2009). Intelligibility of foreign-accented speech in noise for younger and older adults. *Journal of the Acoustical Society of America*, *125*, 2751. <https://doi.org/10.1121/1.4784597>
- Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*, 209–221. <https://doi.org/10.1037/0012-1649.27.2.209>
- Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*, 477–501. <https://doi.org/10.1017/S0305000900010679>, PubMed: 2808569
- Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *Neuroimage*, *186*, 33–42. <https://doi.org/10.1016/j.neuroimage.2018.10.057>, PubMed: 30367953
- Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research*, *38*, 379–412. <https://doi.org/10.1007/S10936-008-9097-8>, PubMed: 19117134
- Folland, N. A., Butler, B. E., Payne, J. E., & Trainor, L. J. (2015). Cortical representations sensitive to the number of perceived auditory objects emerge between 2 and 4 months of age: Electrophysiological evidence. *Journal of Cognitive Neuroscience*, *27*, 1060–1067. https://doi.org/10.1162/jocn_a_00764, PubMed: 25436670
- Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*, 1207–1223. <https://doi.org/10.1080/01690965.2012.701758>
- Foulsham, T., Farley, J., & Kingstone, A. (2013). Mind wandering in sentence reading: Decoupling the link between mind and eye. *Canadian Journal of Experimental Psychology*, *67*, 51–59. <https://doi.org/10.1037/a0030217>, PubMed: 23458551
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138. <https://doi.org/10.1038/nrn2787>, PubMed: 20068583
- Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*, 511–517. <https://doi.org/10.1038/nn.3063>, PubMed: 22426255
- Gordon-Salant, S., Yeni-Komshian, G. H., & Fitzgibbons, P. J. (2010). Recognition of accented English in quiet by younger normal-hearing listeners and older listeners with normal-hearing and hearing loss. *Journal of the Acoustical Society of America*, *128*, 444–455. <https://doi.org/10.1121/1.3397409>, PubMed: 20649238
- Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Schurman, J. (2010). Short-term adaptation to accented English by younger and older adults. *Journal of the Acoustical Society of America*, *128*, EL200–EL204. <https://doi.org/10.1121/1.3486199>, PubMed: 20968326
- Grant, K. W., & Bernstein, J. G. W. (2019). Toward a model of auditory–visual speech intelligibility. In A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Multisensory processes* (pp. 33–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-10461-0_3
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197–1208. <https://doi.org/10.1121/1288668>, PubMed: 11008820
- Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, *53*, 1529–1542. [https://doi.org/10.1044/1092-4388\(2010/09-0005\)](https://doi.org/10.1044/1092-4388(2010/09-0005)), PubMed: 20699342
- Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*, 1333–1341. <https://doi.org/10.3758/S13414-014-0821-1>, PubMed: 25810157
- Haufe, S., Meinecke, F., Görden, K., Dähne, S., Haynes, J. D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96–110. <https://doi.org/10.1016/j.neuroimage.2013.10.067>, PubMed: 24239590
- Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *Neuroimage*, *181*, 617–626. <https://doi.org/10.1016/j.neuroimage.2018.07.052>, PubMed: 30048749
- Hazan, V., Uther, M., & Grunlund, S. (2015). How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles? *18th International Congress of Phonetic Sciences*. <https://wlv.openrepository.com/handle/2436/623993>
- Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850–856. <https://doi.org/10.3758/PBR.16.5.850>, PubMed: 19815788
- Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behaviour. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 20095–20100. <https://doi.org/10.1073/pnas.1213390109>, PubMed: 23151506
- Jang, E. H., Park, B. J., Park, M. S., Kim, S. H., & Sohn, J. H. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, *34*, 25. <https://doi.org/10.1186/s40101-015-0063-5>, PubMed: 26084816
- Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Quarterly Journal of Experimental Psychology*, *65*, 1563–1585. <https://doi.org/10.1080/17470218.2012.658822>, PubMed: 22530648
- 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>, PubMed: 31362048
- Kalashnikova, M., Peter, V., Di Liberto, G. M., Lalor, E. C., & Burnham, D. (2018). Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific Reports*, *8*, 13745. <https://doi.org/10.1038/s41598-018-32150-6>, PubMed: 30214000
- Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English

- consonants produced by non-native speakers. *Journal of the Acoustical Society of America*, *136*, 1352. <https://doi.org/10.1121/1.4892770>, PubMed: 25190408
- Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*, 85–110. https://doi.org/10.1207/S15327078IN0401_5
- Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behaviour and Development*, *24*, 372–392. [https://doi.org/10.1016/S0163-6383\(02\)00086-3](https://doi.org/10.1016/S0163-6383(02)00086-3)
- Knoll, M., & Scharer, L. (2007). Acoustic and affective comparisons of natural and imaginary infant-, foreigner- and adult-directed speech. *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, *3*, 1669–1672. <https://doi.org/10.21437/interspeech.2007-29>
- Knoll, M., Uther, M., & Costall, A. (2011). Using the internet for speech research: An evaluative study examining affect in speech. *Behaviour & Information Technology*, *30*, 845–851. <https://doi.org/10.1080/0144929X.2011.577192>
- Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science*, *17*, 110–124. <https://doi.org/10.1111/DESC.12098>, PubMed: 24176002
- Kothe, C. A. E., & Jung, T. P. (2014). U.S. patent application no. 14/895,440.
- Krasich, K., McManus, R., Hutt, S., Faber, M., D'Mello, S. K., & Brockmole, J. R. (2018). Gaze-based signatures of mind wandering during real-world scene processing. *Journal of Experimental Psychology*, *147*, 1111–1124. <https://doi.org/10.1037/XGE0000411>, PubMed: 29963888
- Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*, 279–292. <https://doi.org/10.1016/j.neuron.2006.12.011>, PubMed: 17224408
- Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*, 110–113. <https://doi.org/10.1126/science.1154735>, PubMed: 18388295
- Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, *42*, 526–539. <https://doi.org/10.1044/jslhr.4203.526>, PubMed: 10391620
- 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, U.S.A.*, *109*, 1431–1436. <https://doi.org/10.1073/pnas.1114783109>, PubMed: 22307596
- Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Frontiers in Psychology*, *7*, 52. <https://doi.org/10.3389/fpsyg.2016.00052>, PubMed: 26869959
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*, 173–190. <https://doi.org/10.2307/3588329>
- Mégevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., et al. (2020). Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex. *Journal of Neuroscience*, *40*, 8530–8542. <https://doi.org/10.1523/JNEUROSCI.0555-20.2020>, PubMed: 33023923
- Mercier, M. R., Molholm, S., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., & Foxe, J. J. (2015). Neuro-oscillatory phase alignment drives speeded multisensory response times: An electro-corticographic investigation. *Journal of Neuroscience*, *35*, 8546–8557. <https://doi.org/10.1523/JNEUROSCI.4527-14.2015>, PubMed: 26041921
- Moradi, S., Lidestam, B., & Rönnerberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy. *Frontiers in Psychology*, *4*, 359. <https://doi.org/10.3389/fpsyg.2013.00359>, PubMed: 23801980
- Moradi, S., Lidestam, B., Saremi, A., & Rönnerberg, J. (2014). Gated auditory speech perception: Effects of listening conditions and cognitive capacity. *Frontiers in Psychology*, *5*, 531. <https://doi.org/10.3389/fpsyg.2014.00531>, PubMed: 24926274
- 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*, 1640–1655. <https://doi.org/10.1037/dev0000750>, PubMed: 31169400
- 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*, 133–137. <https://doi.org/10.1111/J.0963-7214.2004.01502010.X>, PubMed: 14738521
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97. <https://doi.org/10.1111/J.1467-1770.1995.tb00963.x>
- Narayan, C. R., & McDermott, L. C. (2016). Speech rate and pitch characteristics of infant-directed speech: Longitudinal and cross-linguistic observations. *Journal of the Acoustical Society of America*, *139*, 1272–1281. <https://doi.org/10.1121/1.4944634>, PubMed: 27036263
- Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*, 4–12. <https://doi.org/10.1007/S00426-005-0031-5>, PubMed: 16362332
- O'Sullivan, A. E., Lim, C. Y., & Lalor, E. C. (2019). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, *50*, 3282–3295. <https://doi.org/10.1111/ejn.14425>, PubMed: 31013361
- Oostenfeld, 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*, 15689. <https://doi.org/10.1155/2011/156869>, PubMed: 21253357
- Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, *28*, 381–393. <https://doi.org/10.1044/jshr.2803.381>, PubMed: 4046579
- Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, *65*, 553–567. <https://doi.org/10.3758/BF03194582>, PubMed: 12812278
- Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*, e14521. <https://doi.org/10.7554/eLife.14521>, PubMed: 27146891
- 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>, PubMed: 22973251
- Peter, V., Kalashnikova, M., Santos, A., & Burnham, D. (2016). Mature neural responses to infant-directed speech but not adult-directed speech in pre-verbal infants. *Scientific Reports*, 6, 34273. <https://doi.org/10.1038/srep34273>, PubMed: 27677352
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11, 105–110. <https://doi.org/10.1016/j.tics.2006.12.002>, PubMed: 17254833
- Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech Language and Hearing Research*, 52, 1073–1081. [https://doi.org/10.1044/1092-4388\(2009/07-0276\)](https://doi.org/10.1044/1092-4388(2009/07-0276)), PubMed: 19641083
- Porretta, V., Buchanan, L., & Järvikivi, J. (2020). When processing costs impact predictive processing: The case of foreign-accented speech and accent experience. *Attention, Perception, & Psychophysics*, 82, 1558–1565. <https://doi.org/10.3758/S13414-019-01946-7>, PubMed: 31970710
- Porretta, V., & Kyröläinen, A. J. (2019). Influencing the time and space of lexical competition: The effect of gradient foreign accentedness. *Journal of Experimental Psychology: Learning Memory and Cognition*, 45, 1832–1851. <https://doi.org/10.1037/xlm0000674>, PubMed: 30816767
- Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, 58, 1–21. <https://doi.org/10.1016/j.wocn.2016.05.006>
- Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91, 1306–1325. <https://doi.org/10.1109/JPROC.2003.817150>
- Räsänen, O., Kakouros, S., & Soderstrom, M. (2018). Is infant-directed speech interesting because it is surprising?—Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition*, 178, 193–206. <https://doi.org/10.1016/j.cognition.2018.05.015>, PubMed: 29885600
- Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research*, 39, 1159–1170. <https://doi.org/10.1044/jshr.3906.1159>, PubMed: 8959601
- Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, 33, 2329–2337. <https://doi.org/10.1111/J.1460-9568.2011.07685.X>, PubMed: 21615556
- 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, 1147–1153. <https://doi.org/10.1093/cercor/bhl024>, PubMed: 16785256
- Ru, P. (2001). *Multiscale multirate spectro-temporal auditory model* [Unpublished doctoral dissertation]. University of Maryland College Park.
- Rudmann, D. S., McCarley, J. S., & Kramer, A. F. (2003). Bimodal displays improve speech comprehension in environments with multiple speakers. *Human Factors*, 45, 329–336. <https://doi.org/10.1518/hfes.45.2.329.27237>, PubMed: 14529202
- Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, 93, B69–B78. <https://doi.org/10.1016/j.cognition.2004.01.006>, PubMed: 15147940
- 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, e1003743. <https://doi.org/10.1371/journal.pcbi.1003743>, PubMed: 25079216
- Simonetti, S., Kim, J., & Davis, C. (2016). Identifying visual prosody: Where do people look? *Proceedings of the International Conference on Speech Prosody, 2016-January* (pp. 840–844). <https://doi.org/10.21437/SpeechProsody.2016-172>
- Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27, 501–532. <https://doi.org/10.1016/j.dr.2007.06.002>
- Soley, G., & Sebastian-Galles, N. (2020). Infants' expectations about the recipients of infant-directed and adult-directed speech. *Cognition*, 198, 104214. <https://doi.org/10.1016/j.cognition.2020.104214>, PubMed: 32058101
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 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>, PubMed: 35436614
- Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49, 2–7. <https://doi.org/10.1016/j.specom.2006.10.003>
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, U.S.A.*, 102, 1181–1186. <https://doi.org/10.1073/pnas.0408949102>, PubMed: 15647358
- Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30, 555–568. <https://doi.org/10.1006/jpho.2002.0165>
- Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *Journal of Speech, Language, and Hearing Research*, 56, 471–480. [https://doi.org/10.1044/1092-4388\(2012/10-0288\)](https://doi.org/10.1044/1092-4388(2012/10-0288)), PubMed: 23275394
- Zhang, H., Miller, K. F., Sun, X., & Cortina, K. S. (2020). Wandering eyes: Eye movements during mind wandering in video lectures. *Applied Cognitive Psychology*, 34, 449–464. <https://doi.org/10.1002/acp.3632>
- 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, 1417–1426. <https://doi.org/10.1523/JNEUROSCI.3675-12.2013>, PubMed: 23345218