Visual Cortical Tracking of Categorical Speech Features Is Enhanced for Trained Lipreaders

Zhewei Cao

Honors Thesis

Department of Brain and Cognitive Sciences,

University of Rochester, Rochester, NY, USA

Spring, 2020

Abstract

Neuroimaging research has demonstrated that our brain tracks features of observed visual speech during silent lip-reading. Specifically, it has been shown that activity over the occipital scalp can be better decoded when representing speech as a combination of low-level features and categorical speech features, rather than either feature set along, suggesting a tracking of both types of features in visual cortex. However, it remains unclear whether this tracking specifically reflects visible categorical speech features. In addition to visual cortex, it has been shown that auditory cortex is also activated during silent lip-reading, yet not much is known about whether cortical activity is tracking the unheard auditory stimuli in a meaningful way. In the current study, we look into whether silent lip-reading can elicit tracking of categorical visual speech features, as well as the unheard speech envelope in the absence of auditory speech. To do this, we trained participants to be better lip-readers on five audiovisual videos of a speaker, and then tested them on a random selection of the five silent trained videos as well as five silent novel videos of the same speaker while performing a target word detection task. We recorded both behavioral performance and recorded electroencephalography (EEG) data during silent lipreading. Results showed that EEG data from both trained and novel videos saw no clear auditory cortical tracking of the unheard speech signal. However, in visual cortical regions we found enhanced tracking of categorical visual speech features for trained videos over novel videos after regressing out the motion in the videos. Importantly, the extent to which the motion in the videos was tracked by visual regions was the same for both trained and novel videos. With these results, we cast doubt on whether silent lip-reading training does in fact improve cortical representation of the acoustic envelope and found supporting evidence for categorical speech processing at the visual cortex during silent lipreading.

Introduction

Spoken language is a multisensory percept that constitutes the foundation of human interaction. In a face-to-face setting, speech is comprised of not only auditory, but also a visual component. Although the neural processing of auditory speech has been much explored, the visual component of speech however has received much less attention.

Studies have shown that the addition of vision enhances speech perception in compromised hearing conditions (Sumby and Pollack, 1954; Ross et al., 2007) and for populations with impaired hearing (Erber, 1971). This benefit is thought to derive from visual speech interacting with auditory speech through two "modes". The first involves vision providing information about the acoustic auditory signal that are hard to detect (Campbell, 2008). For example, it has been shown that from visual speech, specifically mouth width, listeners are able to extract spectro-temporal information of the auditory speech (Plass, 2019). The second mode involves visual speech taking on a correlated role when visual speech information, such as mouth and jaw movement, are temporally correlated with the acoustic auditory information. This redundancy between information from the two modalities has been found to benefit people with normal hearing in optimal hearing conditions (Reisberg *et al.*, 1987; Jiang *et al.*, 2002).

A few studies have tried to characterize the neurophysiology underlying visual speech alone, i.e., silent lip-reading. It has previously been shown that cortical tracking to visual-only speech was significantly correlated with subject's lip-reading ability (Crosse *et al.*, 2015). To look into which specific components of visual features the subjects are encoding, there has been evidence for the encoding of higher-level visual speech features in visual cortex (Bernstein & Liebenthal, 2014; O'Sullivan *et al.,* 2017). In one study, it was suggested that combining both low-level visual speech features (e.g., frame-to-frame motion) and higher-level visual features (e.g., visemes – defined as visually similar phonemes) more accurately predicts neural activity over occipital scalp. Yet, further evidence is required to definitively prove that the visual cortex indeed engages in categorical speech processing.

Maybe not surprisingly, the auditory cortex might also be playing a role in silent lip-reading -- fMRI literature has suggested the auditory cortex might be activated during silent lipreading (Bernstein *et al.,* 2002; Calvert *et al.*, 1997; Pekkola *et al.*, 2005). A model has suggested that visual speech provides temporal cues about the acoustic signal, which could directly project to and affect the sensitivity and activation of auditory cortex (Calvert *et al.*, 1997; Beauchamp *et al.*, 2004). However, little is known about the quantitative nature of cross-modal activation between the visual and auditory cortex during speech perception, specifically, whether the auditory cortex is being activated in a manner that is meaningfully related to the unheard speech.

Here, to better understand how visual speech can aid auditory speech perception in various hearing conditions, we trained subjects with natural, continuous audiovisual speech stimuli and recorded how their brain represented the unheard speech when they lipread silent visual speech. Specifically, we hypothesized that when subjects could better lipread, the enhanced lipreading ability would be indexed by (1) enhanced tracking of the unheard acoustic speech envelope in the auditory cortex & (2) improved representation of higher-level visual speech features in the visual cortex. After preliminary analysis, we found that neural signals over visual areas better represented a categorical representation of speech (i.e., visemes) when participants were previously trained to lip read that speech, suggesting higher-level, speech-specific processing at the visual cortices during silent lipreading, and further illuminating visual cortex's role in audiovisual speech perception.

Method

Subjects

Sixteen native English speakers (11 females; age range: 19–37 years), none of which were trained lip readers, gave written informed consent to participant in the experiment. All participants were right-handed , had self-reported normal hearing, normal or corrected-to-normal vision, and no underlying neurological disease. The study was approved by the Research Subjects Review Board (RSRB) at the University of Rochester.

Stimuli

The speech stimuli were drawn from a collection of videos featuring a well-known male speaker, whose head, shoulders and chest are centered in the frame. The speech was conversational like, and the linguistic content focused on political policy. These speech stimuli were rendered into 60-s 1280 × 720-pixel movies in VideoPad Video Editor (NCH Software). Soundtracks were deleted from the 15 videos which had a frame rate of 30 frames per second.

Stimulus presentation and data recording took place in a dark, sound-attenuated room with participants seated at a distance of 70 cm from the visual display. Visual stimuli were presented on a 26″ LCD monitor operating at a refresh rate of 60 Hz.

Procedure

Ten different one-minute videos were randomly drawn from the total of 15 videos, five of which selected as training videos. The experiment was divided into two parts: training and testing.

In the first phase, participants were trained to lip-read five randomly selected videos. These videos with intact sound were presented to the participant 10 times each in a random sequence. Subjects were instructed to watch the videos and focus on the lip movements as attentively as they could, while listening to the speech through the provided headphone in the booth. No behavioral testing was carried out at this stage.

After the training phase, we recorded EEG while subjects watched silent versions of the five videos learned during the training phase and five novel videos. To refresh participants' memory, each silent trained video was preceded by the same video with sound. The videos (audiovisual, novel and trained) were randomly placed into a sequence of 15 and were presented to the subjects for a total of 4 iterations during this testing phase. A demonstration of the testing sequence within an iteration could be found in Figure 1.

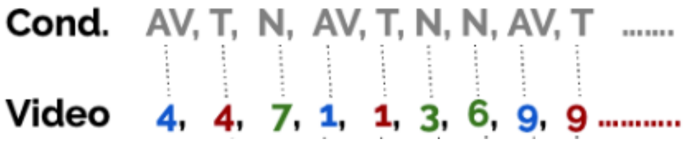

Figure 1, Subject's trial sequence in the testing phase. Subjects have four iterations of 15 trials. Within an iteration, there are five audio-visual trained (AV, blue), five silent trained (T, red) and five silent novel (N, green) videos. Each silent trained video (T) always follows its audiovisual version (AV).

To measure how well participants could lip-read videos, we asked participants to rate their subjective intelligibility of the stimuli on a scale from 0-10 and to perform a word detection task. At the beginning of each video's presentation, subjects were given a target word to detect. This word was unique for each trial and we never reused a target word between trial repetitions. Subjects were instructed to press the space bar on the keyboard when they perceived the occurrence of the target word, e.g. reading lip movements in silent trials, or hearing the word in non-silent trials. At the end of each trail, subjects were asked to rate the perceived intelligibility of the watched video on a scale of 0-10, 0 representing that they could understand 0~10% of the video just watched and 10 being that they could understand 90~100%.

Subjects were instructed to fixate on the speaker's mouth while minimizing eye blinking and all other motor activity.

Calculation of behavioral accuracy (D')

The subjects' behavioral responses were recorded by the Presentation software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA), and were used to calculate the d-prime statistic. The d-prime statistic is considered an ideal measure for how readily the signal could be picked up among noise. (Macmillan and Creelman, 2004). In a signal detection task, it is defined as

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}\left(\sigma_S^2 + \sigma_N^2\right)}} \qquad (1)$$

, with μS and σS signal mean and standard deviation, and μN and σN noise mean and standard deviation. In the case of this experiment, d' serves as a good estimate of the subject's sensitivity to target words in the silent visual speech. Button presses recorded within 2 seconds after target word onset were recorded as hits (signal); and those recorded outside the 2 second timeframe were calculated as false alarms (noise).

EEG Acquisition and Pre-Processing

During the testing phase, continuous EEG data were acquired from subjects using an ActiveTwo system (BioSemi) from 128 scalp electrodes. The data were low pass filtered online below 134 Hz and digitized at a rate of 512 Hz. We synchronized the EEG to the stimulus via triggers that were sent by an Arduino Uno microcontroller that detected an audio click inserted at the beginning of each soundtrack. EEG signal pre-processing was conducted in MATLAB; the data were bandpass filtered between 0.3 Hz and 15 Hz and re-referenced to the average of all channels. To identify channels with excessive noise, the standard deviation of each channel was compared with that of the surrounding channels in MATLAB. Channels which are two standard deviations away from neighboring channels were deemed contaminated by noise and replaced by spline-interpolating the remaining clean channels with weightings based on their relative scalp location in EEGLAB (Delorme and Makeig, 2004). Each trial's data was down sampled to 64Hz before further analysis.

EEG Analysis - Stimulus Feature Extraction

Our overall EEG analysis strategy was based on relating the ongoing dynamics of the recorded EEG to different dynamic features of the audio and visual speech. More particularly, we aimed to fit models of the EEG responses based on different features of the speech, and then to test if those model can successfully predict EEG responses to new stimuli. If some models do a better job than others of predicting EEG, one can then say something about what features of the speech are being reflected in the data and, thus, perhaps, in the brain. A substantial amount of previous research has done this in order to study the hierarchical processing of speech (Di Liberto et al., 2015) and how that processing is affected by attention (Power et al., 2012) and multisensory input (Crosse et al., 2015, 2016).

To test our hypotheses about auditory cortical involvement in silent lip-reading and the possibility that visual cortex might be processing categorical features of visual speech, we represented the speech using several different representations.

Envelopes Here we use the speech envelope as an estimate of the unheard acoustic speech, as demonstrated in previous studies (O'Sullivan *et al.,* 2017). The broadband amplitude envelope representation was obtained by bandpass filtering the speech signal into 256 logarithmically spaced frequency bands between 80 Hz and 3000 Hz using a gammachirp filter bank (Irino and Patterson, 2006).

The envelope at each of the 256 frequency bands was calculated using a Hilbert transform, and the broadband envelope was obtained by averaging over the 256 narrowband envelopes.

Frame-to-frame Motion To accurately capture the motion of pixels in the videos, a frame-to-frame motion was calculated. For each frame, a matrix of motion vectors was calculated using an ‘‘Adaptive Rood Pattern Search’’ block matching algorithm (Barjatya, 2004). Through pooling all motion vector lengths in the frame (Bartels et al., 2008), a global motion vector was obtained. This vector was resampled to 64 Hz to match the sampling rate of the EEG data.

Phoneme/Viseme Groupings of phonemes have been identified to be ambiguous and confusable when presented visually during identification tasks (Woodward and Barber, 1960; Fisher, 1968). Therefore, each of these groups of visually confusable phonemes can be seen as the building blocks of visual speech, i.e. visemes. To derive a viseme representation from our videos, a phonemic representation from Di Liberto et al. (2015) was used. Using methods defined in Auer and Bernstein (1997), the phonemic representation was converted to visemes. The phoneme- to-viseme transformation means that timing of our viseme representation is actually synchronized with the acoustic boundaries rather than the visual.

EEG Analysis - Temporal Response Function Estimation

In order to relate the continuous EEG to the various visual speech representations introduced above, we used the temporal response function (TRF), a regression analysis that enables mapping between EEG and features. A TRF can be thought of as a filter that describes how a particular stimulus feature (e.g., the acoustic envelope) is transformed into the continuous EEG at each channel and is analogous to the event related potential. So if s(t) represents the stimulus feature at time t, the EEG response at channel n, r (t, n), can be modeled as a convolution with a to-be-estimated TRF, w (τ, n).

$$r(t,n) = \sum\nolimits_{\tau = T_{\min}}^{T_{\max}} w(\tau,n)s(t-\tau) + \varepsilon(t,n), \tag{2}$$

where ε(t, n) stands for responses unexplained by the model at each channel. Considering that it takes several tens of milliseconds for the effect of stimulus to be detectable in the responses, the TRF is calculated between stimulus and response across different lengths of time-lags. In the current study, we fit TRFs for each 60-s trial using ridge regression expressed in the following matrix form:

$$w = (S^T S + \lambda I)^{-1} S^T r, \tag{3}$$

where λ is the ridge parameter, chosen to optimize the stimulus- response mapping which also provides regularization and prevents overfitting, S is a matrix containing a time series of stimulus samples for the window of interest (i.e., the lagged time series), *r* is a matrix of the 128-channel neural response data, and *I* is the identity matrix. The TRF was computed using a custom-built toolbox in MATLAB (Crosse *et al.*, 2016).

EEG Prediction and Model Evaluation

With this TRF modeling approach, we set out to assess how some of the abovementioned speech features was being encoded in neural signals using forward prediction, and how some of the univariate speech features were represented in the EEG signals with backward reconstruction. We fit TRFs describing the mapping between the speech features and the EEG. Then, using leave-one-out cross-validation, we examine how well the left-out EEG data could be predicted using the different models. If the EEG data could be predicted with accuracy using a particular model, we can suggest that the EEG is reflecting the encoding of that particular feature or set of features. Because we had 20 trials for each subject, leave-one-out cross-validation meant that each TRF was fit to the data from 19 trials and then the average TRF across these 19 trials was used to predict the EEG in the remaining trial (Crosse *et al.,* 2016; O'Sullivan *et al.,* 2017).

Prediction accuracy was measured by calculating Pearson's (r) linear correlation coefficient between the predicted and original EEG responses at each electrode channel. This procedure is explained in more detail in Crosse et al. (2016). In the case of this study, when two features A and B are related, we attempted to partial out the influence of one from another – For every trial, after we generated predicted EEG from one speech feature A, the predicted signal is subtracted from the true EEG responses. This gives us EEG that has the contribution from feature A partialled out. We then try to predict this residual EEG using feature B. This tells us how well feature B is represented in the EEG after accounting for the contribution of feature A.

Results

Training improves lip-reading ability

All sixteen participants performed a target word detection task while EEG data was acquired from them. To investigate the effect of training on the task performance, we compared the self-reported measure of intelligibility, as well as sensitivity to the target word during the silent lipreading trials.

The training's effect can be measured by both the subject's self-reported measures of intelligibility after each trial, as well as their behavioral performances during the trial. Subjects rated videos that they were trained on as significantly more intelligible (Figure 2a, p = 0.0059, Wilcoxon test). Consistent with their self-report, subjects' performance in target word detection task also improved on trained videos (Figure 2b). The condition had a significant effect on d-prime (p = 0.0076, Wilcoxon test).

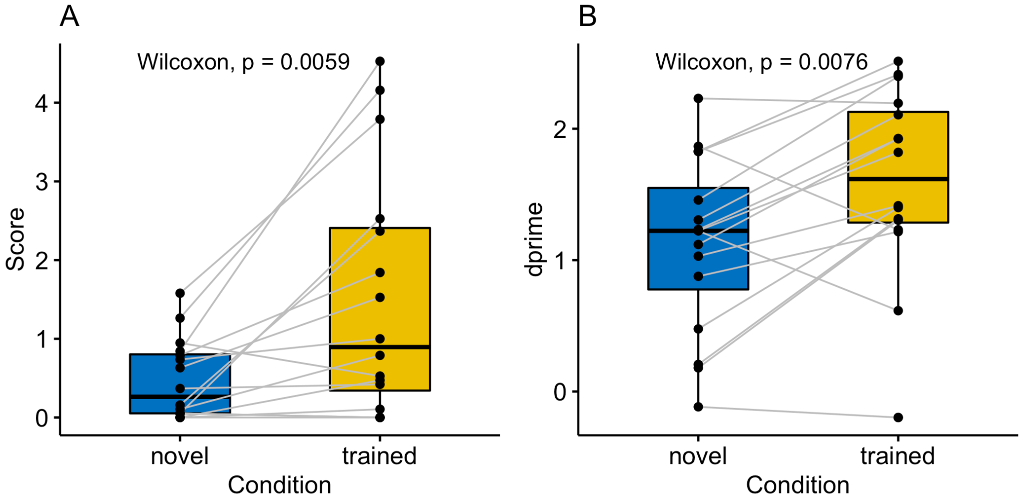

Figure 2, Subject's behavioral responses during the lip-reading task: A) Subject's rating of the silent speech videos for conditions of novel and trained ; B) Subject's performance of target word detection for the silent speech videos for conditions of novel and trained

Unheard Speech Envelope reconstruction is not improved after training

To look into whether the training had an impact on the cortical representation of unheard speech, we reconstructed an estimate of the unheard speech envelope with the collected EEG signal for trained and novel trials (Figure 3a). We found that the unheard envelope could not be more accurately reconstructed within each subject in the trained condition compared with the novel condition (p=0.40, Wilcoxon test). Following this finding, we set out to examine whether regions of the scalp were contributing differently to the two conditions -- we predicted data collected from each EEG channel with the acoustic envelope and visualized the differences between them (Figure 3b). Although there seems to be differences between the two conditions in the frontal central scalp, signals collected from the frontal-central channels found no difference between novel and trained conditions (Figure 3c).

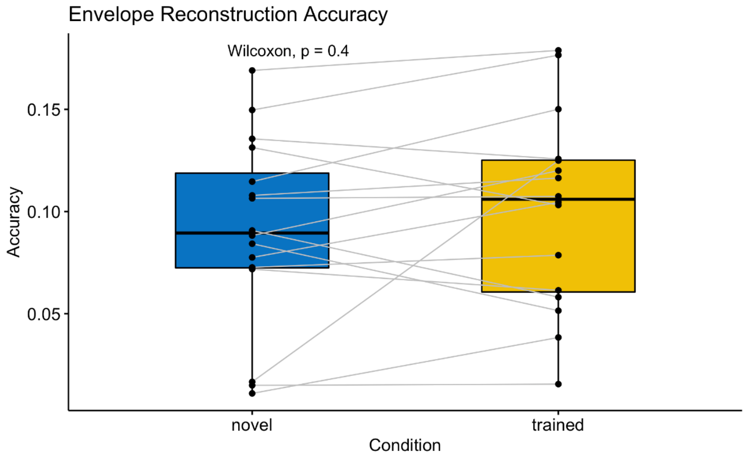

Figure 3a, the reconstruction accuracies of unheard speech envelope for novel and trained conditions (n=16, p=0.40, Wilcoxon)

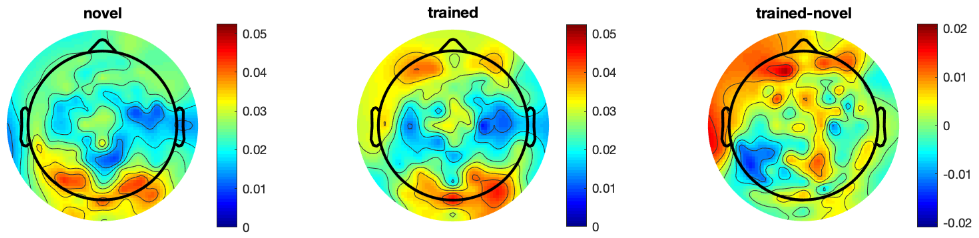

Figure 3b Topographic maps of prediction accuracies using the unheard speech envelope

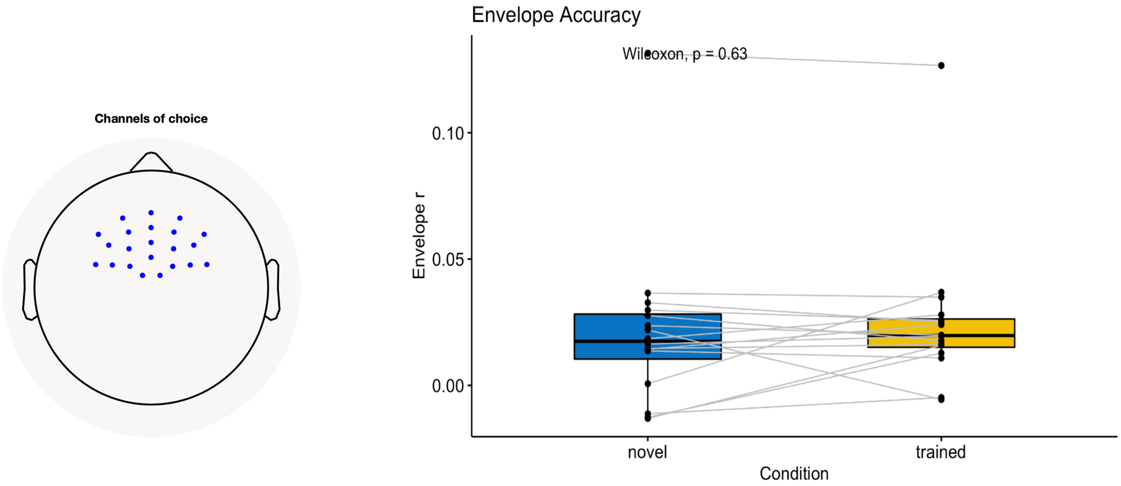

Figure 3c Left: 22 channels chosen over the frontal-central region

Right: Mean prediction accuracy accuracies using the acoustic speech envelope over the chosen channels for each subject in both novel and trained conditions

Encoding of Frame-to-Frame Motion

The frame to frame motion stimulus used in the current study includes both local and global motion in the visual scene (Bartels et al., 2008, O'Sullivan et al., 2017). While some of the global and local motion components may be tangential to speech (e.g., movement of speaker relative to the background), some of them are related (e.g. jaw movements). In this way, the frame-to-frame vector provides an estimate of general low-level visual information received by the visual cortices.

To look into whether training has an impact on the cortical responses to the silent speech videos, we first examined how the frame-to-frame motion is encoded in both trained and novel conditions (Figure 4a). We showed an enhancement in motion tracking for trained vs novel conditions particularly in the visual regions. However, to account for the fact that motion and visemes are correlated with each other (Files *et al.*, 2015), and to examine frame to frame motion's unique contribution to visual speech perception, we first regressed out visemes, and then related the residual EEG to the motion vectors (Figure 4b). We then chose 25 channels at the occipital area and found no difference between trained and novel conditions (Figure 4c, p=0.38, Wilcoxon test).

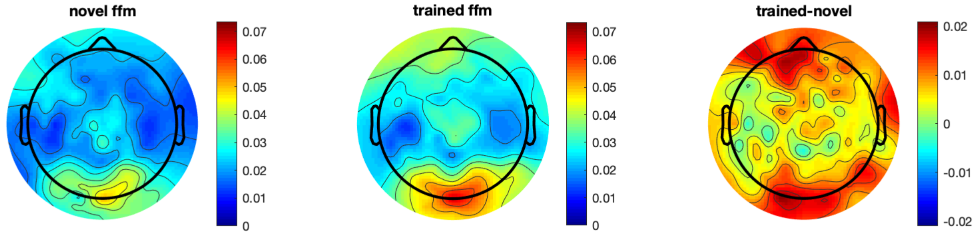

Figure 4a Topographic maps of prediction accuracies using frame to frame motion in the unheard speech

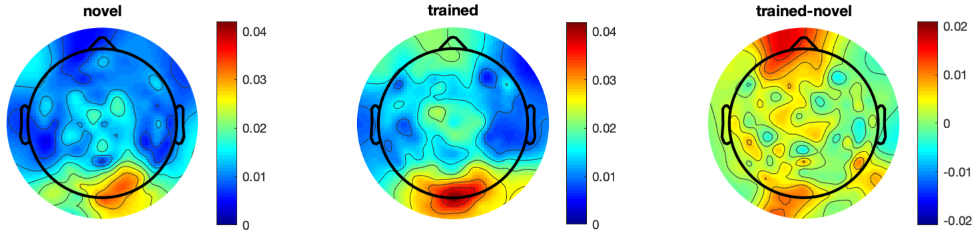

Figure 4b Topographic maps of prediction accuracies using frame to frame motion in the unheard speech after regressing out potential contributions from visual phonetic features

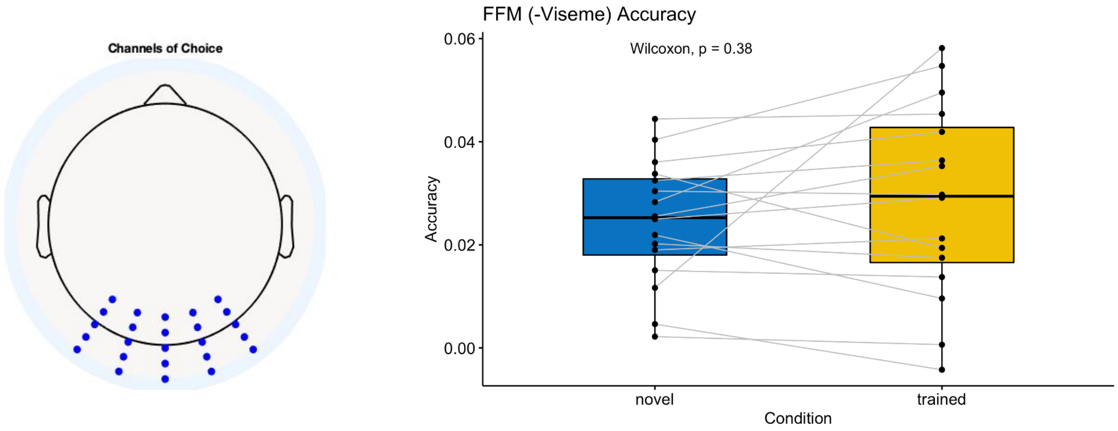

Figure 4c Left:25 channels chosen over the occipital region

Right: Mean prediction accuracy accuracies using frame-to-frame motion while excluding visual phonetic features over the chosen channels for each subject in both novel and trained conditions

Encoding of Visemes

It has been suggested that EEG may be reflecting categorical speech processing beyond encoding the onset and offset of speech, yet more evidence is needed for whether speech-specific, categorical-viseme level processing were ongoing at the visual cortices. We reasoned that, if such processing can happen in visual areas, it should be specifically enhanced during successful lipreading. To test this, we wanted to examine how training participants to lipread certain videos would affect the encoding of visemes. In both auditory (Di Liberto et al., 2015) and visual speech (O'Sullivan et al., 2017; Hauswald et al., 2018), the relationship between EEG and phoneme representations of speech has been explored.

We found that the higher prediction accuracies for visemes in the visual regions for trained conditions (Figure 5a). Again, to assess the unique contribution of viseme to visual speech perception, we regressed out frame-to-frame motion and related the residual EEG data to the viseme vectors (Figure 5b). To quantify the enhanced tracking of visemes only at the occipital area, we examined 25 channels over the occipital region and found significant improvement in viseme representation for trained videos compared with novel (Figure 5c, p=0.0092, Wilcoxon test).

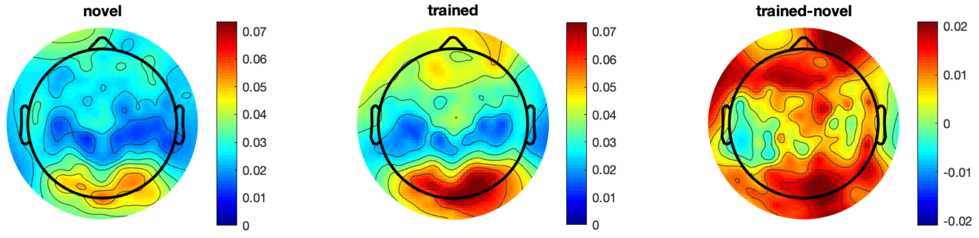

Figure 5a Topographic maps of prediction accuracies using visual phonetic features (i.e. visemes) in the unheard speech

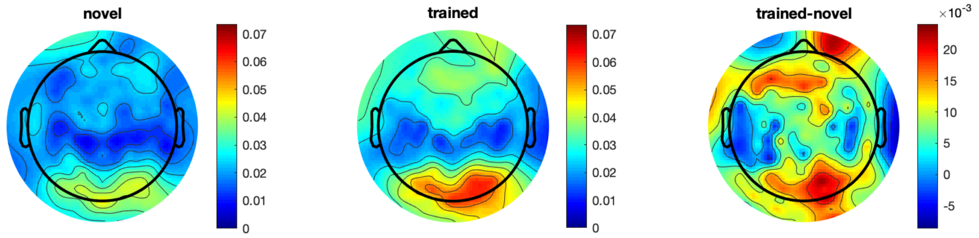

Figure 5b Topographic maps of prediction accuracies using visual phonetic features (i.e. visemes) in the unheard speech, after regressing out the influence from low level visual features (i.e. frame-to-frame motion).

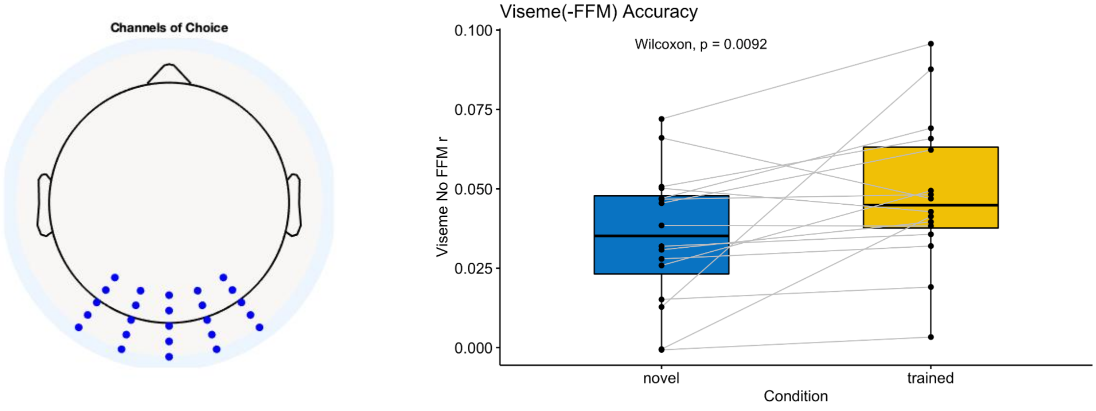

Figure 5c Left:25 channels chosen over the occipital region

Right: Mean prediction accuracy accuracies using visual phonetic features (i.e. visemes) while excluding frame-to-frame motion over the chosen channels for each subject in both novel and trained conditions

Discussion

In the current study, we trained subjects to lipread with audiovisual videos. Our results show that this training was successful. We demonstrate that during lipreading of trained videos, neural signals are representing categorical representations of speech more robustly. This suggests that the visual cortex is engaged in processing of categorical visual speech features, besides encoding of physical stimulus dynamics.

Training didn't improve tracking of the unheard speech envelope

Our finding that cortical representation of the unheard speech envelope wasn't enhanced when subjects were able to better lipread (Figure 3) seems to be at odds with previous findings that showed cortical tracking of unheard acoustic information (Bourguignon et al. 2020). Bourguignon et al. showed that auditory cortices are entraining to frequencies below 1 Hz, which matches with rhythmicity of phrases and sentences. However, they failed to observe a similar tracking for frequency between 4-8 Hz, the syllable rate in the speech. Contrary to previous findings of enhanced tracking for frequency between 4-8 Hz in the auditory cortex for intelligible auditory speech than unintelligible (Ahissar et al., 2001; Luo and Poeppel, 2007; Peelle et al., 2013), the current study like Bourguignon and colleagues, also failed to find similar differences in tracking of silent visual speech between more intelligible (trained) and less intelligible (novel) video conditions.

Representation of categorical visual speech features is enhanced with improved lipreading abilities

We found no difference in tracking of low-level visual features in the occipital area between trained and novel conditions, suggesting that this said tracking is not speech specific. This finding is consistent with previous work in the field(Hauswald *et al.* 2018). Hauswald and colleagues found no difference in tracking of low-level visual features when silent visual speech videos were played forward and backward, also implying the non-speech-specific nature of tracking of motion in the visual cortices.

We also found supporting evidence for the encoding of categorical speech-specific visual features in the occipital area. This is also in line with previous work, which has also shown that EEG signals in the occipital area can be better decoded with a model with categorical visual speech features than one without (O'Sullivan et al., 2017). Further, other studies have identified posterior superior temporal sulcus (pSTS) and the posterior medial temporal gyrus (MTG) as loci for visuo-phonological processing (Bernstein *et al.*,2011), and preliminary results from the current study supports this hypothesis.

Limitations and Future Directions

It is important to consider some limitations of the current study. First, the training paradigm of the current study might not be entirely effective due to the length of the training stimuli. For the normal hearing population, 60 seconds of audiovisual stimuli is a lot to memorize and lipreading is a difficult task after all. Second, in the vein of poor training outcomes, for the target word detection task, some subjects might be remembering visual markers of the target word, instead of truly learning to lipread. This would've led to a higher sensitivity to the target word despite showing similar enhancement in cortical tracking. Importantly, however, if this was true, it would only decrease the effect size of any visual speech processing. As such, our increased measure of viseme processing for trained speech is likely a lower bound.

The current study could benefit from a number of further statistical analysis on the current data. First of all, it'd be interesting to explore cortical representations of specific frequency-bands of the unheard auditory stimuli to answer questions like whether the auditory cortex is better at tracking <1 Hz frequency during trained trials. Second of all, it would also be of interest to examine individual differences in the collected data. Such analysis could shed light on the relationships between behavioral measures, e.g. improvement in intelligibility and that in lipreading abilities; as well as correlation between behavioral and neurophysiological measures, e.g. improvement in lipreading abilities and the higher entrainment to stimuli in the cortex.

Acknowledgements & Contributions

This research was supported by National Institutes of Health Grant RO1 DC016297.

Aisling O’Sullivan M.S., Lauren A. Szymula M.A., Aaron R. Nidiffer, PhD and Edmund C. Lalor, PhD helped complete the project. The experiment was designed and implemented by ZC under the supervision of AOS, AR and EL. The experiment data was collected by ZC and LS , and analyzed by ZC, AOS and AR. The manuscript was written by ZC and reviewed and edited by AOS, AR and EL.

The authors report no conflicts of interest.

Reference

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci U S A* 98:13367–13372.

Barjatya, A. (2004). Block matching algorithms for motion estimation. *IEEE Trans. Evol. Comput*. 8, 225–239.

Bartels, A., Zeki, S., and Logothetis, N. K. (2008). Natural vision reveals regional specialization to local motion and to contrast-Invariant, global flow in the human brain. *Cereb. Cortex* 18, 705–717. doi: 10.1093/cercor/ bhm107

Bernstein, L. E., Auer, E. T. J., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport*, 13(3), 311–315.

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

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus. *Neuron*, 41(5), 809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *Journal of Neuroscience*, *40*(5), 1053–1065.

Calvert, G. A., dullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., Mcguire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–596. doi: 10.1126/science.276.5312.593

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb. 2007.2155

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5:e1000436.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195-14204.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10:604. doi: 10.3389/fnhum.2016.00604

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003. 10.009

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *J. Speech Hear. Res.* 14, 496–512. doi: 10.1044/jshr.1403.496

Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.00878

Fisher, C. G. (1968). Confusions among visually perceived consonants. *J. Speech Hear. Res*. 11, 796–804. doi: 10.1044/jshr.1104.796

Goncalves, N. R., Whelan, R., Foxe, J. J., and Lalor, E. C. (2014). Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: a general linear modeling approach to EEG. *Neuroimage* 97, 196–205. doi: 10. 1016/j.neuroimage.2014.04.012

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. *Current Biology: CB*, 28(9), 1453-1459.e3. https://doi.org/10.1016/j.cub.2018.03.044

Jiang J, Alwan A, Keating PA, Auer ET, Bernstein LE (2002) On the relationship between face movements, tongue movements, and speech acoustics. EURASIP J Appl Signal Process 11:1174 –1188

Lazard DS, Giraud A-L (2017) Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. *Nat Commun* 8:14872.

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54:1001–1010.

Macmillan, N. A., & Creelman, C. D. (2004). Detection theory: A user's guide. Psychology press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Frontiers in Human Neuroscience*, 10, 679. doi:10.3389/fnhum.2016.00679

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23:1378–1387.

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

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport* 16, 125–128. doi: 10.1097/00001756-200502080-00010

Plass, J., Brang, D., Suzuki, S., & Grabowecky, M. (2019, May 20). Vision Perceptually Restores Auditory Spectral Dynamics in Speech.

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur J Neurosci* 35:1497–1503.

Reisberg D, McLean J, Goldfield A (1987) Easy to hear but hard to under- stand: a lip-reading advantage with intact auditory stimuli. In: Hearing by eye: the psychology of lip-reading (Dodd B, Campbell R, eds), pp 97–114. Hillsdale, NJ: Erlbaum.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex*, *17*(5), 1147-1153.

Simpson, A. J.; Fitter, M. J. (1973). "What is the best index of detectability?". Psychological Bulletin. 80 (6): 481–488.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acous. Soc. Am*. 26, 212–215. doi: 10.1121/1.1907309

Woodward, M. F., and Barber, C. G. (1960). Phoneme perception in lipreading. *J. Speech Hear. Res*. 3, 212–222. doi: 10.1044/jshr.0303.212