7th Annual International IEEE EMBS Conference on Neural Engineering
Montpellier, France, 22 - 24 April, 2015

Investigating the Temporal Dynamics of Auditory Cortical Activation to Silent Lipreading*

Michael J. Crosse, Hesham A. ElShafei, John J. Foxe and Edmund C. Lalor, *Member, IEEE*

Abstract— **Neuroimaging research has demonstrated that observing visual speech in the absence of auditory speech activates primary auditory cortex. However, it remains unclear what this activation precisely reflects. It is well established that, during continuous auditory speech, neural activity in auditory cortex tracks the temporal envelope of the speech signal. Recently, it has been suggested that this process may in fact reflect an internal synthesis of the speech stream rather than the encoding of the envelope per se. Could silent lipreading therefore elicit a similar "entrainment" to the envelope in the absence of auditory speech? Here, we test this hypothesis by examining the impact of lipreading accuracy on envelope tracking using electroencephalography (EEG). We provide evidence to suggest that the EEG response over left temporal scalp tracks the unheard speech more faithfully during accurate lipreading. We also demonstrate that the envelope can be reconstructed from EEG data recorded during silent lipreading with accuracy above chance level. This could have implications for brain-computer interface technology.**

I. Introduction

Viewing a speaker's lip movements (lipreading) greatly enhances speech perception [1]. Neuroimaging research has sought to identify the stage of processing at which visual and auditory information interacts. Using functional magnetic resonance imaging (fMRI), it has been demonstrated that silent lipreading activates primary auditory cortex [2]. While fMRI has furthered our understanding of multisensory integration, it is not well suited to examining the nature of rapidly modulating cortical activity over time. This task is better suited to methods with higher temporal resolution such as electro- and magneto-encephalography (EEG/MEG). Such techniques have reliably shown that auditory cortical activity tracks the envelope of acoustic speech [3]. Recently, Ding et al. [4] demonstrated that this process may instead reflect an analysis-by-synthesis mechanism, whereby speech features that are correlated with the envelope are encoded during the synthesis phase, thus leading to envelope tracking. Many visual cues involved in lipreading are also correlated with the acoustic envelope [5]. Encoding of such features could therefore manifest in a process that also aligns with the speech envelope. Here, we test this hypothesis by examining the impact of lipreading accuracy on the entrainment of EEG to the unheard speech.

Work by O'Sullivan et al. [6] has demonstrated that it is possible to reconstruct an estimate of the speech envelope from EEG data. While this work has distinct applications in brain-computer interface (BCI) technology, such methods would better serve BCIs by decoding the users' inner thoughts, i.e., covert speech. Such an approach presents two main challenges: (1) how do we model the neural representation of an internal process, and (2) how do we determine the exact time at which it occurred? Recently, Martin et al. [7] successfully decoded covert speech from electrocorticography (ECoG) data using a decoder that modelled the neural representation of overt speech, while timing issues were dealt with using dynamic time warping. In this study, we demonstrate how the natural statistics of visual speech can be utilized to overcome these issues: (1) assuming that speech perception and imagery share a partially overlapping cortical representation, the original acoustic signal can be used as an estimate of what the perceiver imagined, and (2) timing issues are naturally circumvented because the perceiver is continually prompted to imagine the speech time-locked to the visual stimulus. Here, we apply the method of stimulus reconstruction as a quantitative measure of envelope tracking in EEG during silent lipreading.

II. Methods

A. Subjects

Twelve native English speakers (5 females; age range: 22–37 years), none of which were trained lipreaders, gave written informed consent. All subjects were right-handed, free of neurological diseases, had normal hearing and normal or corrected-to-normal vision. The experiment was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.

B. Stimuli and Procedure

The stimuli were drawn from a collection of videos featuring a male speaker reciting fluent American English sentences. Fifteen 60-s videos were rendered into 1280 × 720-pixel movies at a fame rate of 30 FPS in VideoPad Video Editor (NCH Software). Soundtracks were deleted from 14 of the 15 videos used for silent lipreading. The remaining video was preserved in audiovisual (AV) format and used as a control. The soundtrack was sampled at 48 kHz with 16-bit resolution. It was compressed in Audacity Audio Editor to amplify lower intensities and thus boost the signal-to-noise ratio (SNR) of the neural response during testing.

*Research supported by the Irish Higher Education Authority's Graduate Research Education Programme in Engineering.

M. J. Crosse is with the School of Engineering and Trinity Centre for Bioengineering, Trinity College Dublin, Dublin 2, Ireland (e-mail: crossemj@tcd.ie).

H. A. ElShafei was with Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland. He is now with the Lyon Neuroscience Research Center, University Lyon, Lyon, France (e-mail: hesham.elshafei@inserm.fr).

J. J. Foxe is with the Department of Pediatrics & Neuroscience, Albert Einstein College of Medicine, Bronx, USA (e-mail: john.foxe@einstein.yu.edu).

E. C. Lalor is with the School of Engineering, Trinity Centre for Bioengineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland (phone: +353-1-8961743; e-mail: edlalor@tcd.ie).

Stimulus presentation was controlled using software by Presentation (Neurobehavioral Systems) and delivered using a 19-inch CRT monitor and Sennheiser HD650 headphones. Prior to EEG testing, subjects were trained on the AV stimulus to ensure familiarity with the speech content. During EEG testing, the same AV stimulus was presented 14 times as a control. This known video (Vk) was also presented in visual-only format 14 times, for which subjects were instructed to lipread. The remaining 14 unknown videos (Vu) were presented once each in visual-only format. Subjects were instructed to lipread the Vu stimuli even though they were not familiar with the audio content. Stimulus presentation order was randomized across conditions within subjects. During each 60-s trial, subjects were required to respond to a target word with a button press. A different set of target words was used for each condition and the assignment of target words was counterbalanced across subjects. Each target word occurred between 1 and 3 times per trial and there were 28 targets in total per condition.

C. EEG Acquisition and Pre-processing

EEG was recorded at 130 locations (128 scalp and left and right mastoids) using an ActiveTwo system (BioSemi). Triggers indicating the start of each trial were sent using an Arduino Uno microcontroller which detected an audio click at the start of each soundtrack. The data were low-pass filtered online below 134 Hz and digitized at a rate of 512 Hz. Subsequent pre-processing was conducted offline in MATLAB (MathWorks); the data were band-pass filtered between 1 and 25 Hz and re-referenced to the average of the mastoid channels. The time series were visually inspected in Cartool (brainmapping.unige.ch/cartool) to identify channels with excessive noise. Channels contaminated by noise were recalculated by spline-interpolating the surrounding clean channels in EEGLAB [8].

D. Temporal Response Function Estimation

To examine the relationship between the neural response and the presented stimulus, we calculated the temporal response function (TRF) for each of the three conditions [3]. A TRF can be interpreted as a filter, W, that describes the brain's linear transformation of the speech envelope to the continuous neural response at each channel location. Speech envelopes were extracted using a Hilbert Transform, filtered below 25 Hz and downsampled to 512 Hz. For each 60-s trial, TRFs were calculated between time lags from −100 to 400 ms using the following ridge regression:

$$W = \left(S^{\mathrm{T}}S + \lambda M\right)^{-1} S^{\mathrm{T}}R \qquad (1)$$

where S is a matrix of the lagged time series of the speech envelope, R is a matrix of all 128 channels of neural response data, M is the regularization term used to prevent overfitting and λ is the ridge parameter, empirically chosen to preserve component amplitude (see [3] for further details).

E. Stimulus Reconstruction

To obtain a cortical measure of stimulus encoding, we determined the fidelity with which we could reconstruct the speech envelope from the neural data [9]. For each 60-s trial, we calculated the decoder, G, which represents the mapping from the neural response at all channel back to the stimulus. The data were downsampled to 64 Hz and the decoders were fit with time lags from −500 to 0 ms as follows:

$$G = \left(R^{\mathrm{T}}R + \lambda I\right)^{-1} R^{\mathrm{T}}S \qquad (2)$$

where R is a matrix of the lagged EEG data, S is the speech envelope and I is the identity matrix. For each subject, a leave-one-out cross validation was performed to reconstruct an estimate of each of the 14 stimuli per condition. Specifically, each estimate was obtained by convolving the neural data corresponding to the test trial with the average of the 13 decoders allocated for training as follows:

$$\hat{S} = R\overline{G} \qquad (3)$$

where $\hat{S}$ is the estimated speech envelope, $\overline{G}$ is the mean of the training decoders and R is the lagged test EEG data. Reconstruction accuracy was measured by performing a Pearson's correlation on the estimated and original envelopes. For each subject, we conducted a separate search of λ over the range 2^{10}, 2^{11},…, 2^{30} such that it optimized reconstruction accuracy within each condition. The λ-value with the highest mean reconstruction accuracy over the 14 trials was chosen to prevent overfitting. However, in the AV and Vk conditions, the same stimulus was repeated over the 14 trials which may have caused overfitting. An additional analysis was thus included which removed any potential bias by averaging the decoders across subjects, within trials [6].

All statistical analyses were conducted using one-way repeated measures ANOVAs and Greenhouse-Geisser correction was applied where necessary. Post hoc comparisons were made using two-tailed paired t-tests, except where otherwise stated.

III. Results

A. Behavior

Twelve subjects performed a target detection task during EEG recording. Reaction times (RTs) were measured from the onset of auditory voicing and hits were counted for responses that were made 200–2000 ms after target onset. Condition had a significant impact on both hit rate ($F_{(2,22)}$ = 76.2, $p < 0.001$; Fig. 1*A*) and RT ($F_{(1.3,14.7)}$ = 24.2, $p < 0.001$; Fig. 1*B*). Planned comparisons showed that subjects were significantly more accurate for Vk (74 ± 11%, mean ± SD) compared to Vu (33 ± 15%; $t_{(11)}$ = 9.2, $p < 0.001$) and that RTs were faster for Vk (532 ± 123 ms) relative to Vu (787 ± 150 ms; $t_{(11)}$ = 7.0, $p < 0.001$).

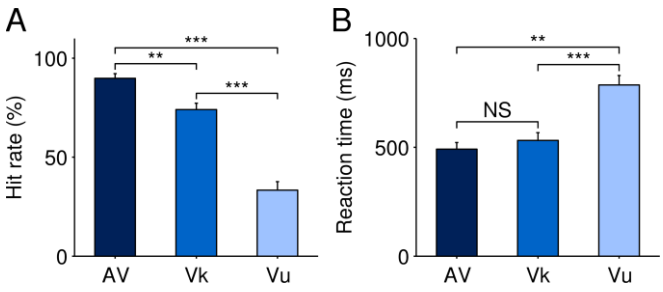

Figure 1. Behavioral performance. *A*, Mean hit rates for AV, Vk and Vu speech. *B*, Mean reaction times for all three conditions. Error bars indicate SEM across subjects. Brackets indicate pairwise statisitcal comparisons (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, NS = not significant).

B. *Spatiotemporal Representation of Speech*

The EEG TRF [3] contains two major response components: a negativity at ~80 ms ($N1_{TRF}$) and positivity at ~130 ms ($P2_{TRF}$; Fig. 2*A*). Fig. 2*B* shows the topography of the $N1_{TRF}$ (left) and $P2_{TRF}$ (right) components. TRF SNR was defined as 0 to 250 ms (signal) and −100 to 0 ms (noise). Over fronto-temporal scalp, SNR was significantly lower for Vk (1.3 ± 1 dB, mean ± SEM) and Vu (0.45 ± 1 dB) compared to AV (6.5 ± 1.9 dB; $F_{(2,22)} = 4.9$, $p < 0.05$; Fig. 2*A*, top) but was similar for all three over occipital scalp ($F_{(2,22)} = 0.07$, $p > 0.05$; Fig. 2*A*, bottom). This is reflected in the statistical cluster maps (Fig. 2*C*) which show significant activation across subjects between 100–200 ms over parieto-occipital scalp in all three conditions and also over fronto-temporal scalp in the AV condition (Fig. 2*C*, top; $p < 0.05$).

To compare the responses of the visual conditions (Vk and Vu) to those of the control condition (AV), a series of Pearson's correlations were performed on their TRFs (0–250 ms). Fig. 3 shows the correlation coefficient (r) at each channel location. Channels where r is significantly greater than zero across subjects are indicated by black markers ($p < 0.05$). The comparison between Vk and Vu revealed a significant cluster of channels over occipital scalp (Fig. 3, right). Interestingly, there was also a cluster over left temporal scalp in the AV and Vk comparison (Fig. 3, left).

C. *Cortical Encoding of Speech*

Stimulus reconstruction was applied using two different approaches. In the first approach, decoders were averaged across trials, within subjects and conditions [9]. We found that condition had a significant impact on reconstruction accuracy ($F_{(2,22)} = 29.5$, $p < 0.001$; Fig. 4*A*). A planned comparison showed that reconstruction accuracy for Vk (0.1 ± 0.03, mean ± SD) was significantly higher than that for Vu (0.08 ± 0.03; $t_{(11)} = 2.5$, $p < 0.05$). Although care was taken to optimize regularization within each condition, it remains a possibility that the conditions with repeated stimuli (AV and Vk) were somewhat biased. In the second analysis, this bias was removed by averaging the decoders across subjects, within trials and conditions [6]. The impact of condition on reconstruction accuracy was weakened by this approach ($F_{(2,22)} = 6.2$, $p < 0.01$; Fig. 4*B*). There was also no significant difference in reconstruction accuracy between Vk (0.041 ± 0.02) and Vu (0.044 ± 0.02; $t_{(11)} = 0.5$, $p > 0.05$). While mean reconstruction accuracy values were significantly reduced across all three conditions, they were still higher than the 95th percentile of chance level (Fig. 4*B*).

IV. Discussion

Here, we tested the hypothesis that auditory cortex synthesizes the unheard speech signal during silent lipreading and that this synthesis is reflected in the neural tracking of the speech envelope. Specifically, we demonstrated that the temporal profile of the neural response to silent lipreading was significantly correlated with that of audiovisual speech over left temporal scalp, but only when lipreading was accurately perceived. We also showed that an estimate of the acoustic envelope could be reconstructed from EEG data recorded during silent lipreading with accuracy greater than chance level.

The temporal response function (TRF), which maps sensory input to cortical activation, was used as a direct measure of envelope tracking [3]. We found that, although TRF SNR was relatively low over fronto-temporal scalp during silent lipreading (Fig. 2*A*, top), its temporal profile was significantly correlated with that of audiovisual speech when lipreading was accurately perceived (Fig. 3, left). This may suggest that accurate processing of visual speech features plays a role in envelope tracking, in line with work espousing an analysis-by-synthesis mechanism [4]. This is also supported by numerous studies that have reported attentional effects on envelope tracking (e.g., [10]). Indeed, we must consider the possibility that using the same stimulus in two of the three conditions may have had an impact on the similarity of the TRFs. In theory, this should not influence the correlation between TRFs because a TRF represents the

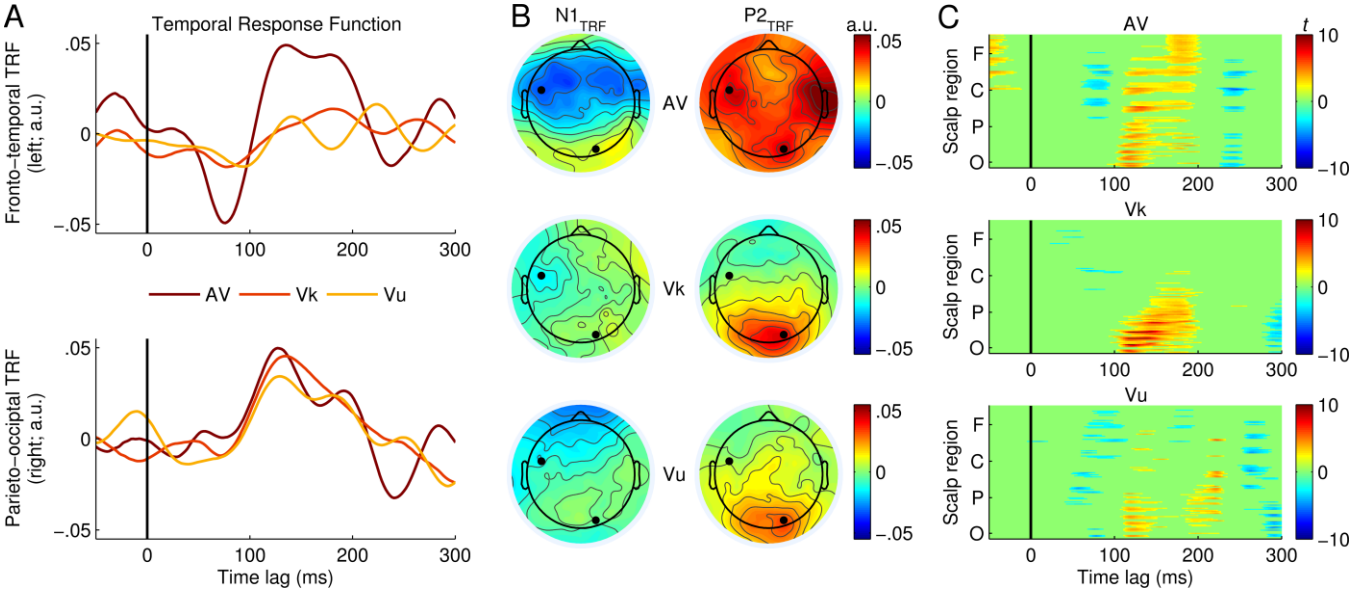

Figure 2. Temproal response function (TRF) timecourse and topography. *A*, TRFs over left fronto-temporal scalp (top) and right parieto-occipital scalp (bottom). *B*, Topographies of $N1_{TRF}$ components (left) and $P2_{TRF}$ components (right). Black markers indicate channel locations plotted in *A*. *C*, Statisitical cluster maps show where and when TRF amplitude is significantly different to zero ($p < 0.05$, t-tests; F = frontal, C = central, P = parietal, O = occipital).

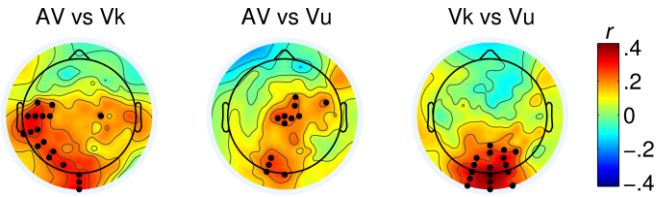

Figure 3. Mean correlation coefficients (r) between the TRFs (0–250 ms) of each condition at each channel location. Black markers indicate channels where r is significantly greater than zero across subjects ($p < 0.05$, t-tests).

impulse response to a unit change in stimulus intensity [3]. This is supported by the fact that the TRFs over occipital scalp were very similar in the Vk and Vu conditions (Fig. 3, right), even though different visual stimuli were presented.

During silent lipreading, auditory cortex is not directly stimulated via the auditory nerve, hence TRF SNR over fronto-temporal scalp was reduced in the Vk and Vu conditions relative to the AV condition (Fig. 2*A*, top). This issue could be addressed by presenting audio noise that is spectrally matched to the absent speech signal [4]. This would directly stimulate the auditory nerve which may help boost auditory cortical responses. There was no difference in SNR over occipital scalp (Fig. 2*A*, bottom) because each condition was matched in terms of visual stimulus intensity. The regression analysis is sensitive to this occipital activation because instantaneous measures of motion during visual speech are highly correlated with the amplitude of the acoustic envelope [10]. However, in keeping with an analysis-by-synthesis mechanism, this occipital activity may in fact reflect the processing of visual speech features in visual cortex as opposed to just motion tracking. It has been shown that every level of speech structure can be perceived visually, thus suggesting that there are visual modality-specific representations of speech in visual brain areas and not just in auditory brain areas (for a review, see [11]).

The method of stimulus reconstruction was used as an alternative measure of envelope tracking [9]. We found that the fidelity with which we could reconstruct an estimate of the unheard acoustic signal was significantly improved when the subject could accurately lipread (Fig. 4*A*). This was true when each decoder was optimized separately within each condition so as not to bias those with repeated stimuli. However, as stated earlier, we cannot be certain that the AV and Vk decoders were not somewhat biased. Future work will address this issue by assigning a different stimulus to each condition (counterbalanced across subjects) and presenting each one an equal number of times to ensure equal bias. In the absence of such experimental manipulations, a within-trial analysis was carried out which removed any potential bias from the AV and Vk conditions. However, because this approach involved averaging decoders over subjects, the decoders were grossly generalized and reconstruction accuracies dropped considerably (Fig. 4*B*). This is caused by the inherent spatiotemporal variability in the neural activity of the twelve subjects [6].

V. Conclusion

In summary, we have presented evidence to suggest that accurate lipreading may elicit envelope tracking in auditory

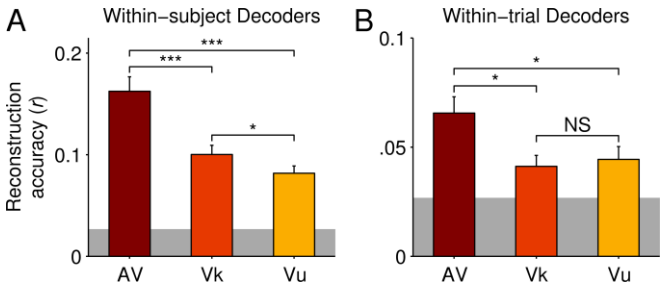

Figure 4. Reconstruction accuracy. *A*, Mean reconstruction accuracy of decoders fit within subjects, across trials. *B*, Mean reconstruction accuracy of decoders fit within trials, across subjects. The shaded area represents the 95th percentile of chance level (permutation test). Error bars indicate SEM across subjects. Brackets indicate pairwise statisitcal comparisons (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, NS = not significant).

cortex reflective of visual speech processing. Although we cannot conclude on the impact of lipreading accuracy on stimulus encoding, we have demonstrated that it is possible to reconstruct an estimate of the envelope of covert speech from EEG data by utilizing the natural statistics of visual speech. While similar results have been demonstrated using ECoG [7], we suggest that EEG may provide a non-invasive and cost-effective solution to decoding imagined thoughts for future BCI technology. This may have implications for future visual speech research, as well as non-invasive BCI methodologies. As outlined above, future work will implement a modified paradigm to extend the results presented here.

References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.,* vol. 26, pp. 212-215, 1954.

[2] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. R. Williams, P. K. McGuire*, et al.*, "Activation of auditory cortex during silent lipreading," *Science,* vol. 276, pp. 593-596, Apr 1997.

[3] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *Eur. J. Neurosci.,* vol. 31, pp. 189-193, Jan 2010.

[4] N. Ding, M. Chatterjee, and J. Z. Simon, "Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure," *Neuroimage,* 2013.

[5] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," *PLoS Comput. Biol.,* vol. 5, Jul 2009.

[6] J. A. O'Sullivan, M. J. Crosse, A. J. Power, and E. C. Lalor, "The effects of attention and visual input on the representation of natural speech in EEG," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 2800-2803.

[7] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger*, et al.*, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Front. Neuroeng.,* vol. 7, 2014.

[8] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods,* vol. 134, pp. 9-21, Mar 2004.

[9] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex," *J. Neurophysiol.,* vol. 102, pp. 3329-3339, Dec 2009.

[10] A. J. Power, J. J. Foxe, E. J. Forde, R. B. Reilly, and E. C. Lalor, "At what time is the cocktail party? A late locus of selective attention to natural speech," *Eur. J. Neurosci.,* vol. 35, pp. 1497-1503, May 2012.

[11] L. E. Bernstein and E. Liebenthal, "Neural pathways for visual speech perception," *Front. Neurosci.,* vol. 8, 2014.