Systems/Circuits

Neurophysiological Indices of Audiovisual Speech Processing Reveal a Hierarchy of Multisensory Integration Effects

Aisling E. O'Sullivan,[1] Michael J. Crosse,[2] Giovanni M. Di Liberto,[3] Alain de Cheveigné,[3,4] and Edmund C. Lalor[1,5]

[1]School of Engineering, Trinity Centre for Biomedical Engineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, [2]X, The Moonshot Factory, Mountain View, CA and Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York 10461, [3]Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, Paris Sciences et Lettres University, Centre National de la Recherche Scientifique, Paris 75005, France, [4]University College London Ear Institute, University College London, London WC1X 8EE, United Kingdom, and [5]Department of Biomedical Engineering and Department of Neuroscience, University of Rochester, Rochester, New York 14627

Seeing a speaker's face benefits speech comprehension, especially in challenging listening conditions. This perceptual benefit is thought to stem from the neural integration of visual and auditory speech at multiple stages of processing, whereby movement of a speaker's face provides temporal cues to auditory cortex, and articulatory information from the speaker's mouth can aid recognizing specific linguistic units (e.g., phonemes, syllables). However, it remains unclear how the integration of these cues varies as a function of listening conditions. Here, we sought to provide insight on these questions by examining EEG responses in humans (males and females) to natural audiovisual (AV), audio, and visual speech in quiet and in noise. We represented our speech stimuli in terms of their spectrograms and their phonetic features and then quantified the strength of the encoding of those features in the EEG using canonical correlation analysis (CCA). The encoding of both spectrotemporal and phonetic features was shown to be more robust in AV speech responses than what would have been expected from the summation of the audio and visual speech responses, suggesting that multisensory integration occurs at both spectrotemporal and phonetic stages of speech processing. We also found evidence to suggest that the integration effects may change with listening conditions; however, this was an exploratory analysis and future work will be required to examine this effect using a within-subject design. These findings demonstrate that integration of audio and visual speech occurs at multiple stages along the speech processing hierarchy.

Key words: **CCA; EEG; hierarchical processing; multisensory integration; speech in noise; speech in quiet**

Significance Statement

During conversation, visual cues impact our perception of speech. Integration of auditory and visual speech is thought to occur at multiple stages of speech processing and vary flexibly depending on the listening conditions. Here, we examine audiovisual (AV) integration at two stages of speech processing using the speech spectrogram and a phonetic representation, and test how AV integration adapts to degraded listening conditions. We find significant integration at both of these stages regardless of listening conditions. These findings reveal neural indices of multisensory interactions at different stages of processing and provide support for the multistage integration framework.

Received Apr. 18, 2020; revised Mar. 16, 2021; accepted Mar. 22, 2021.

Author contributions: A.E.O., G.M.D.L., A.d.C., and E.C.L. designed research; A.E.O. and M.J.C. performed research; A.d.C. contributed unpublished reagents/analytic tools; A.E.O. analyzed data; A.E.O., M.J.C., G.M.D.L., A.d.C., and E.C.L. wrote the paper.

This work was supported by the Science Foundation Ireland Career Development Award 15/CDA/3316 and by National Institutes of Health National Institute on Deafness and Other Communication Disorders Grant R01 DC016297.

The authors declare no competing financial interests.

*Correspondence should be addressed to Edmund C. Lalor at edmund_lalor@urmc.rochester.edu.

https://doi.org/10.1523/JNEUROSCI.0906-20.2021

Copyright © 2021 the authors

Introduction

One prominent theory of speech perception is that speech is processed in a series of computational steps that follow a hierarchal structure, with different cortical regions being specialised for processing different speech features (Scott and Johnsrude, 2003; Hickok and Poeppel, 2007; DeWitt and Rauschecker, 2012). One key question is how visual input influences processing within this hierarchy.

Behavioral studies have shown that seeing the face of a speaker improves speech comprehension (Sumby and Pollack,

1954; Grant and Seitz, 2000; Ross et al., 2007). This behavioral advantage is thought to derive from two concurrent processing modes: a correlated mode, whereby visual speech dynamics provide information on auditory speech dynamics, and a complementary mode, where visual speech provides information on the articulatory patterns generating the auditory speech (Campbell, 2008). It seems plausible that the information provided by these two modes would influence levels of the auditory hierarchy differently. Indeed, this idea aligns well with a growing body of evidence indicating that audiovisual (AV) speech integration likely occurs over multiple stages (Schwartz et al., 2004; van Wassenhove et al., 2005; Eskelund et al., 2011; Baart et al., 2014; Peelle and Sommers, 2015). One recent perspective (Peelle and Sommers, 2015) suggests that these stages could include an early stage, where visual speech provides temporal cues about the acoustic signal (correlated mode), and a later stage, where visual cues that convey place and manner of articulation could be integrated with acoustic information to constrain lexical selection (complementary mode). Such early-stage integration could be mediated by direct projections from visual cortex that dynamically affect the sensitivity of auditory cortex (Calvert et al., 1997; Grant and Seitz, 2000; Tye-Murray et al., 2011; Okada et al., 2013), whereas for later-stage integration, articulatory visual cues could be combined with acoustic information in supramodal regions such as the superior temporal sulcus (STS; Beauchamp et al., 2004; Kayser and Logothetis, 2009; Zhu and Beauchamp, 2017; Karas et al., 2019).

While the evidence supporting multiple stages of AV speech integration is compelling, there are several ways in which this multistage model needs to be further developed. First, much of the supporting evidence has been based on experiments involving simple (and often illusory) syllabic stimuli or short segments of speech. This has been very valuable, but it also seems insufficient to fully explore how a correlated mode of AV integration might derive from dynamic visual cues impacting auditory cortical processing. Testing the model with natural speech will be necessary (Theunissen et al., 2000; Hamilton and Huth, 2018). Second, directly indexing neurophysiological representations of different acoustic and articulatory features will be important for validating and further refining the key idea that integration happens at different stages. And third, it will be important to test the hypothesis that this multistage model is flexible, whereby the relative strength of integration effects at different stages might depend on the listening conditions and the availability of visual information.

These are the goals of the present manuscript. In particular, we aim to build on recent work that examined how visual speech affected neural indices of audio speech dynamics using naturalistic stimuli (Luo et al., 2010; Zion Golumbic et al., 2013; Crosse et al., 2015a, 2016b). We aim to do so by incorporating ideas from recent research showing that EEG and MEG are sensitive not just to the acoustics of speech, but also to the processing of speech at the level of phonemes (Di Liberto et al., 2015; Khalighinejad et al., 2017; Brodbeck et al., 2018). This will allow us to derive indices of dynamic natural speech processing at different hierarchical levels and to test the idea that AV speech integration occurs at these different levels, in line with the multistage model (Peelle and Sommers, 2015). Finally, we also aim to test the hypothesis that, in the presence of background noise, there will be a relative increase in the strength of AV integration effects in EEG measures of phoneme-level encoding, reflecting an increased reliance on articulatory information when speech is noisy. To do all this, we introduce a new framework for indexing the electrophysiology of AV speech integration based on canonical correlation analysis (CCA).

Materials and Methods

The EEG data analyzed here were collected as part of previous studies published by Crosse et al. (2015a, 2016b).

Participants

Twenty-one native English speakers (eight females; age range: 19–37 years) participated in the speech in quiet experiment. Twenty-one different participants (six females; age range: 21–35) took part in the speech in noise experiment. Written informed consent was obtained from each participant beforehand. All participants were native English speakers, were free of neurologic diseases, had self-reported normal hearing, and had normal or corrected-to-normal vision. The experiment was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.

Stimuli and procedure

The speech stimuli were drawn from a collection of videos featuring a trained male speaker. The videos consisted of the speaker's head, shoulders, and chest, centered in the frame. The speech was conversational-like and continuous, with no prolonged pauses between sentences. Fifteen 60-s videos were rendered into 1280×720-pixel movies in VideoPad Video Editor (NCH Software). Each video had a frame rate of 30 frames per second, and the soundtracks were sampled at 48 kHz with 16-bit resolution. The intensity of each soundtrack, measured by root mean square, was normalized in MATLAB (MathWorks). For the speech in noise experiment, the soundtracks were additionally mixed with spectrally matched stationary noise to ensure consistent masking across stimuli (Ding and Simon, 2013; Ding et al., 2014) with signal-to-noise ratio (SNR) of −9 dB. The noise stimuli were generated in MATLAB using a 50th-order forward linear predictive model estimated from the original speech recording. Prediction order was calculated based on the sampling rate of the soundtracks (Parsons, 1987).

In both experiments, 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 19-inch CRT monitor operating at a refresh rate of 60 Hz. Audio stimuli were presented diotically through Sennheiser HD650 headphones at a comfortable level of ~65 dB. Stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). For the speech in quiet experiment each of the 15 speech passages was presented seven times, each time as part of a different experimental condition. Presentation order was randomized across conditions, within participants. While the original experiment had seven conditions, here we focus only on three conditions audio-only (A), visual-only (V), and congruent AV (AV). For the speech in noise experiment, however, there were only three conditions (A, V, and AV) and so the passages were ordered 1–15 and presented three times with the condition from trial-to-trial randomized. This was to ensure that each speech passage could not be repeated in another modality within 15 trials of the preceding one. Participants were instructed to fixate on either the speaker's mouth (V, AVc) or a gray crosshair (A) and to minimize eye blinking and all other motor activity during recording.

For both experiments participants were required to respond to target words via button press. Before each trial, a target word was displayed on the monitor until the participant was ready to begin. All target words were detectable in the auditory modality except during the V condition, where they were only visually detectable. A target word was deemed to have been correctly detected if subjects responded by button press within 0–2 s after target word onset. In addition to detecting target words, participants in the speech-in-noise experiment were required to rate subjectively the intelligibility of the speech stimuli at the end of each 60-s trial. Intelligibility was rated as a percentage of the total words understood using a 10-point scale (0–10%, 10–20%, ... 90–100%).

EEG acquisition and preprocessing

The EEG data were recorded using an ActiveTwo system (BioSemi) from 128 scalp electrodes and two mastoid electrodes. The data were low-pass filtered on-line below 134 Hz and digitized at a rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. Subsequent preprocessing was conducted off-line in MATLAB; the data were detrended by subtracting a 50th-order polynomial fit using a robust detrending routine (de Cheveigné and Arzounian, 2018). The data were then bandpass filtered using second-order, zero phase-shift Butterworth filters between 0.3 and 30 Hz, downsampled to 64 Hz, and rereferenced to the average of the mastoid channels. Excessively noisy channels were detected using the spectrogram, kurtosis and probability methods provided by the EEGLAB toolbox. Channels were marked for rejection using a threshold value of three for each method. The channels marked for rejection were interpolated from the surrounding clean channels using the spline function from EEGLAB (Delorme and Makeig, 2004).

Indexing neurophysiological speech processing at different hierarchical levels

Because our aim was to examine how visual information affects the neural processing of auditory speech at different hierarchical levels, we needed to derive separable EEG indices of processing at these levels. To do this, we followed work from Di Liberto et al. (2015), who modeled EEG responses to speech in terms of different representations of that speech. Specifically, they showed that EEG responses to speech were better predicted using a representation of speech that combined both its low-level acoustics (i.e., its spectrogram) and a categorical representation of its phonetic features. The underlying idea is that EEG responses might reflect the activity of neuronal populations in auditory cortex that are sensitive to spectrotemporal acoustic fluctuations and of neuronal populations in association cortices (e.g., the superior temporal gyrus) that may be invariant to spectrotemporal differences between utterances of the same phoneme and, instead, are sensitive to that phoneme category itself. As such, for the present study, we calculated two different representations of the acoustic speech signal.

Spectrogram

This was obtained by first filtering the speech stimulus into 16 frequency bands between 80 and 3000 Hz using a compressive gammachirp auditory filter bank that models the auditory periphery. The gammachirp toolbox was obtained by direct request to the corresponding author on the paper (Irino and Patterson, 2006). Then the amplitude envelope for each frequency band was calculated using the Hilbert transform, resulting in 16 narrow band envelopes forming the spectrogram representation.

Phonetic features

This representation was computed using the Prosodylab-Aligner (Gorman et al., 2011), which, given a speech file and the corresponding textual orthographical transcription, automatically partitions each word into phonemes from the American English International Phonetic Alphabet (IPA) and performs forced alignment (Yuan and Liberman, 2008), returning the starting and ending time points for each phoneme. Manual checking of the alignment was then carried out and any errors corrected. This information was then converted into a multivariate time series that formed a binary array, where there is a one representing the onset and duration of each phoneme and zeros everywhere else. To describe the articulatory and acoustic properties of each phoneme a 19-dimensional phonetic feature representation was formed using the mapping defined previously (Chomsky and Halle, 1968; Mesgarani et al., 2014). This involves mapping each phoneme (e.g., /b/) into a set of phonetic features (e.g., bilabial, plosive, voiced, obstruent) and results in a phonetic feature matrix of ones and zeros that is of dimension 19 (which is the number of phonetic features) by time.

CCA

We wished to see how these different speech representations might be reflected in EEG activity. Previous related research has relied on a regression-based approach that aims to reconstruct an estimate of some univariate feature of the speech stimulus (e.g., its amplitude envelope) from multivariate EEG responses (Crosse et al., 2015a, 2016b). However, because we have multivariate speech representations, we sought to use a method based on CCA (Hotelling, 1936; de Cheveigné et al., 2018), which was implemented using the NoiseTools toolbox (http://audition.ens.fr/adc/NoiseTools/).

CCA works by rotating two given sets of multidimensional data into a common space in which they are maximally correlated. This linear transformation is based on finding a set of basis vectors for each of the given datasets such that the correlation between the variables, when they are projected on these basis vectors, is mutually maximized. In our case, our two datasets are the multidimensional stimulus representation, $X(t)$, of size $T \times J_1$, where T is time and J_1 is the number of features in that representation ($J_1 = 16$ frequency bands of a spectrogram, or $J_1 = 19$ phonetic features), and an EEG data matrix $Y(t)$ of size $T \times J_2$, where T is time and $J_2 = nx\tau$, where n is the number of EEG channels (128) and τ is the number of time lags. The reason for using multiple time lags is to allow for the fact that a change in the stimulus impacts the EEG at several subsequent time lags. In our analysis we included time lags from 0–500 ms, which at a sampling rate of 64 Hz resulted in 32 time lags. For these two data matrices, CCA produces transform matrices A and B of sizes $J_1 \times J_0$ and $J_2 \times J_0$, respectively, where J_0 is at most equal to the smaller of J_1 and J_2. The optimization problem for CCA is formulated as a generalized eigenproblem with the objective function:

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \rho^2 \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix},$$

where C_{XY} is the covariance of the two datasets X and Y and C_{XX} and C_{YY} are the autocovariances, and ρ is the components correlation. Ridge regularization can be performed on the neural data to prevent overfitting in CCA as follows (Vinod, 1976; Leurgans et al., 1993; Cruz-Cano and Lee, 2014; Bilenko and Gallant, 2016):

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \rho^2 \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} + \lambda I \end{pmatrix}.$$

Only the EEG by time-lags matrix (C_{YY}) is regularized (over the range $\lambda = 1 \times 10^{-2}$, 1×10^{-1}..., 1×10^{4}, 5×10^{4}, 1×10^{5}, 5×10^{5}, 1×10^{6}, 5×10^{6}) since the stimulus representations are of low-dimensionality (16 and 19 dimensions). The rotation matrices (A and B) are learned on all trials except one and are then applied to the left-out data which produces canonical components (CCs) for both the stimulus representation and the EEG using the following equation:

$$X_{J1}(t)\mathrm{A} \rightarrow CC_{stim} \rightarrow \rho \leftarrow CC_{resp} \leftarrow Y_{J2}(t)B.$$

The regularization parameter was chosen as the value that gave the highest correlation value (between the stimulus and EEG components) for the average of all the test trials from the leave-one-out cross-validation procedure for each CC. Therefore, the optimal regularization value could vary across components. The rotation weights A and B are trained to find what stimulus features influence the EEG and what aspects of the EEG are responsive to the stimulus, respectively, to maximize the correlation between the two multivariate signals. When the rotation weights are applied to the left-out data we get the CCs of the stimulus (CC_{stim}) and of the response data (CC_{resp}). The first pair of CCs define the linear combinations of each data set with the highest possible correlation. The next pair of CCs are the most highly correlated combinations orthogonal to the first, and so-on (de Cheveigné et al., 2018).

Indexing multisensory integration using CCA

We wished to use CCA to identify any neural indices of multisensory integration during the AV condition beyond what might be expected from the unisensory processing of audio and visual speech. We sought to do this by modeling the encoding of the speech representations in the A and V EEG data and then investigating whether there is some

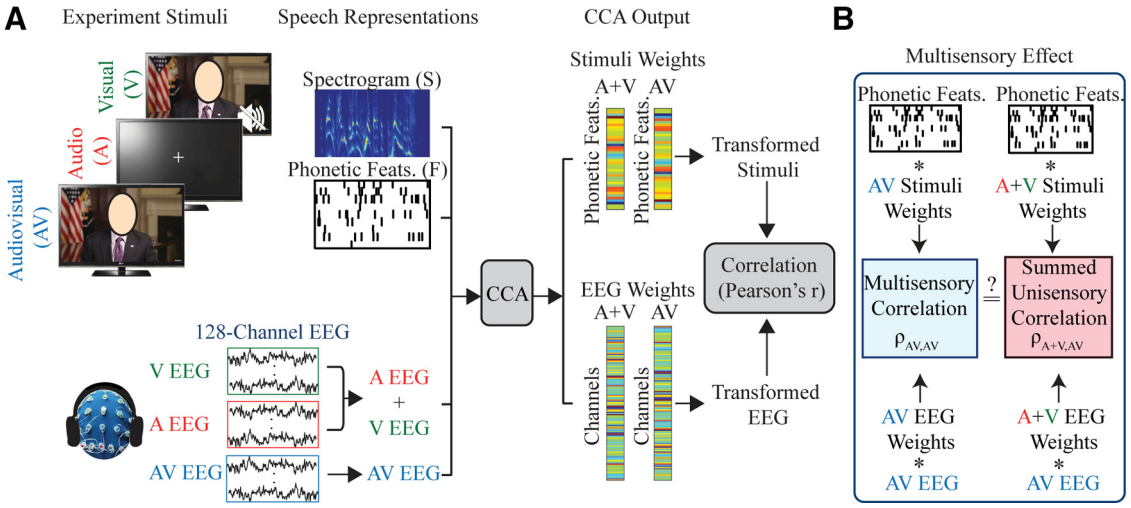

Figure 1. Experiment set-up and analysis approach. The face of the speaker is blocked with an oval for publication but was not blocked for the experiment. ***A***, The stimulus representations used are the spectrogram and the phonetic features and are estimated directly from the speech stimuli. Below, EEG recordings corresponding to each condition. The unisensory A and V EEG are summed to form an A + V EEG data set. The EEG and speech representations are used as inputs to the CCA to determine the optimum weights for rotating the EEG and the given stimulus representation for maximizing the correlation between the two. ***B***, The model built using the A + V data is then tested on the left-out AV data to determine the presence of a multisensory effect.

difference in the speech-related activity in the AV EEG data which is not present in either of the unisensory conditions. In other words, and in line with a long history of multisensory research (Berman, 1961; Stein and Meredith, 1993; Klucharev et al., 2003; van Wassenhove et al., 2005; Besle et al., 2008), we sought to compare AV EEG responses to A + V EEG responses using CCA and to attribute any difference [i.e., AV – (A + V)] to multisensory processing.

To implement this, we summed the EEG data from matching A and V stimuli (i.e., A and V stimuli that came from the same original AV video; Fig. 1*A*). Thus, for each of the original 15 videos, we ended up with AV EEG responses and corresponding A + V EEG responses. Then, we used CCA to relate the multivariate speech representations (spectrogram + phonetic features) to each of these two EEG responses (AV and A + V). This provides two sets of rotation matrices, one between the stimulus and the AV EEG and one between the stimulus and the A + V EEG.

Now, if the scalp recorded EEG activity for the AV condition is simply the auditory and visual modalities being processed separately with no integration occurring, then the A + V and AV EEG responses should be essentially identical. And we would then expect the rotation matrices learned on the A + V EEG data to be identical to those learned on the AV EEG data. Carrying this logic even further, we would then expect to see no differences in the canonical correlation values obtained from the AV data when using the CCA rotation matrices found by training on the A + V EEG data compared with the matrices found by training on the AV EEG data (Fig. 1*B*). In other words, we compared the correlation values obtained when we applied the AV weights (i.e., the A and B matrices found by training on AV EEG data) to left-out AV data, with the correlation values obtained when applying the A + V weights (i.e., the A and B matrices found by training on A + V EEG data) to the AV data (Fig. 1*B*). If there is some difference in the EEG response dynamics for multisensory (AV) compared with the summed unisensory activity (A + V) then we would expect this to have a significant effect on the canonical correlations since the A + V weights would not capture this whereas the AV weights would. To measure the size of this difference we calculated multisensory gain using the following equation:

$$MSI_{Gain} = \frac{\rho_{AV,AV} - \rho_{A+V,AV}}{|\rho_{AV,AV}| + |\rho_{A+V,AV}|},$$

where ρ is the CCs correlation, the first subscript represents the rotations used and the second subscript represents the data on which those rotations are applied (Fig. 1*B*). The difference in performance between the models is normalized since the cortical tracking for very noisy speech is typically much weaker than the tracking of clean speech (Ding and Simon, 2013; Crosse et al., 2016b) and cortical tracking correlation values can also vary substantially across subjects because of differences in cortical folding, skull thickness and scalp thickness. Thus, normalizing the difference in correlations ensures that results from all subjects in both conditions are represented such that they can be compared fairly. One caveat of this normalization approach is that saturation effects (values tending toward ±1) can occur when two inputs have different signs, thus it is important not to assume a normal distribution for this measure.

Statistical analysis

All statistical comparisons were conducted using non-parametric permutation with 10,000 repetitions such that no assumptions were made about the sampling distribution (Combrisson and Jerbi, 2015). This was done by randomly assigning the values from the two groups being compared (pairwise for the paired tests, and non-pairwise for the unpaired tests) and calculating the difference between the groups. This process was repeated 10,000 times to form a null distribution of the group difference. Then the tail of this empirical distribution is used to calculate the p value for the actual data, and two-tailed tests are used throughout. Where multiple comparisons were conducted p values were corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). All numerical values are reported as mean ± SD.

Results

Robust indices of multisensory integration for the speech spectrogram and phonetic features

To investigate the encoding of more complex multivariate representations of the speech stimulus and to isolate measures of multisensory integration at different levels of the speech processing hierarchy, we performed CCA on the AV EEG data using the spectrogram and phonetic features, having trained the CCA on (different) AV data and A + V data. We first sought to do this separately for the spectrogram representation and the phonetic features representation to see whether using either or both of these representations might show evidence of multisensory integration. And we also sought to do this for both our clean speech and noisy speech datasets.

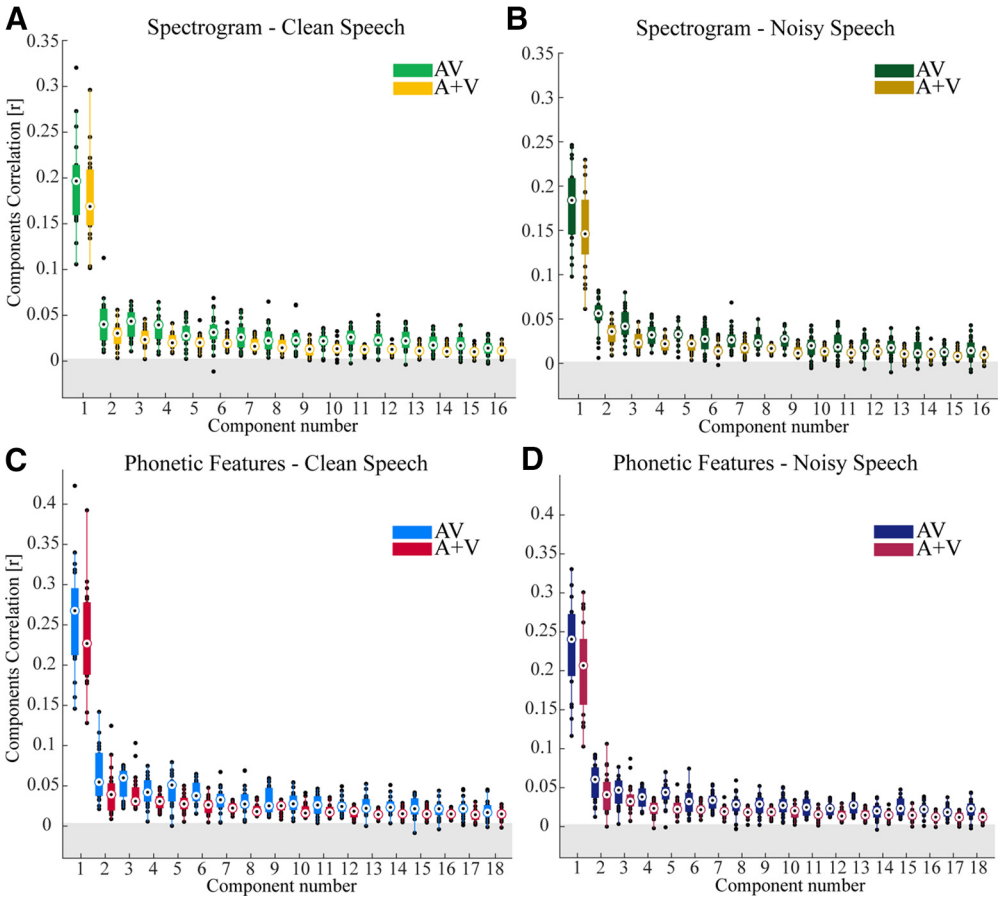

Figure 2. CCA analysis using the spectrogram and phonetic feature representation of the speech stimulus. ***A***, ***B***, The canonical correlations for the spectrogram representation for speech in quiet and speech in noise, respectively. ***C***, ***D***, Canonical correlations for the phonetic feature representation for speech in quiet and speech in noise, respectively. The gray band represents an approximate chance level. All AV and A + V components performed above chance level $p < 0.0001$ for speech in quiet, and for speech in noise $p < 0.0001$. The boxplot function with compact format from MATLAB was used here. The edges of the boxplots display the 25th and 75th percentiles, and the median is marked by a black dot inside a white circle. A data point is determined to be an outlier if it is greater than $Q_{75th} + w \times (Q_{75th} - Q_{25th})$ or less than $Q_{25th} - w \times (Q_{75th} - Q_{25th})$, where w is the maximum whisker length and Q_{25th} and Q_{75th} are the 25th and 75th percentiles of the data.

In both conditions (clean and noisy speech) and for both representations (spectrogram and phonetic features), the correlations for the first component were significantly higher than for all other components. This suggests that the first component captures a substantial percentage of the influence of the speech on the EEG data. And, importantly, both representations showed evidence of multisensory integration.

For the spectrogram representation, we found significant multisensory effects (AV > A + V) for the first CC for speech in quiet ($p < 0.0001$) and the first CC for speech in noise ($p < 0.0001$, FDR corrected p values; Fig. 2*A*,*B*). Indeed, we found multisensory effects for 15/16 components for speech in quiet and for 15/16 components for speech in noise.

A similar pattern was observed when examining the stimulus-EEG relationship using the phonetic feature representation of speech. Specifically, we also found multisensory effects for the first component for clean speech ($p < 0.0001$) and for speech in noise ($p < 0.0001$, FDR corrected). And we found multisensory effects for 13/18 components for speech in quiet and for all 18 components for speech in noise. Although there are 19 phonetic features, there are only 18 components since CCA cuts of eigenvalues below a threshold, which in this case was 10^{-12}, and the last component of phonetic features did not survive this cutoff.

The fact that the spectrogram and phonetic feature representations produced qualitatively similar patterns of multisensory integration was not surprising. This is because both representations are mutually redundant; a particular phoneme will have a characteristic spectrotemporal signature. Indeed, if each utterance of a phoneme were spoken in precisely the same way every time, then the spectrogram and phonetic feature representations would be functionally identical (Di Liberto et al., 2015). But, as we discuss below, in natural speech different utterances of a particular phoneme will have different spectrograms. So, to identify the unique contribution of "higher-level" neurons that are invariant to these spectrotemporal differences and are sensitive to the categorical phonemic features we will need to index the EEG responses that are uniquely explained by the phonetic feature representation while controlling for the spectrogram representation (see below, Isolating multisensory effects at the spectrotemporal and phonetic levels).

Spatiotemporal analysis of CCs: increased cross-modal temporal integration and possible increased role for visual cortex for speech in noise

The previous section showed clear evidence of multisensory integration in the component correlation values obtained from CCA. But how can we further investigate these CCA components to better understand the neurophysiological effects underlying these numbers? One way is to examine how these multisensory effects might vary as a function of the time lag between the

stimulus and EEG and how any effects at different time lags might be represented across the scalp. This is very much analogous to examining the spatiotemporal characteristics of event-related potentials (ERPs) with EEG.

To investigate the spatiotemporal properties of the AV and A + V CCA models we ran the CCA at individual time lags from −1 to 1.5 s. We chose to focus our analysis on the first three CCs. This was mostly to allow investigation of the dominant first component, but also to check whether or not useful insights might be gleaned from any of the subsequent components. For the first component of the spectrogram model, there was significant differences between AV and A + V at −300 to −200, 0–125, and at 500–750 ms for speech in quiet (FDR corrected). For speech in noise there was significant differences at time shifts of −650 to −250 and −60–750 ms (Fig. 3*A*,*D*, FDR corrected). For the second and third components there was no clear pattern to the single time-lag correlation as it was quite flat across all time shifts for both clean and noisy speech (Fig. 3*B*,*C*,*E*,*F*).

For the phonetic feature representation we found significant differences between AV and A + V at −370 to −125 and 90–750 ms for speech in quiet and for speech in noise, there was significant differences at time shifts of −300 to −125 and 300–750 ms (Fig. 3*G*,*J*, FDR corrected). For the second component the pattern reflected that of an onset response and while there was no difference between AV and A + V for speech in quiet there was a small window of difference for speech in noise at 0–75 ms (Fig. 3*H*,*K*, FDR corrected). There was no clear pattern to the single time-lag correlation for the third component for both clean and noisy speech (Fig. 3*I*,*L*). In general, the number of lags at which there was a significant difference between AV and A + V was greater for noisy speech than clean speech, which is consistent with findings in Crosse et al. (2016b). The finding of significant differences at negative lags may be because of the fact that the EEG data are time locked to the onset of the audio and since the visual information often precedes the audio (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014), the AV data may contain some information about the speech at "negative" time lags. Another possibility, however, is that the effect at negative lags is because of the autocorrelation of the stimulus and the autocorrelation of the EEG. Nonetheless, in our multilag CCA we have only used positive lags (0–500 ms) and so any effects at negative lags will not influence our overall results.

In summary, the single lag analysis reveals multisensory interactions for both stimulus representations at similar ranges of lags, mostly in the 0- to 500-ms range. It also shows the dominance of the first component in capturing the relationship between the EEG and stimulus, however for phonetic features the second component also appears to display a time-locked response. Nonetheless, there are many fewer time lags for which we see multisensory interactions in the second and third components. The first component for both spectrogram and phonetic features shows an early peak which is likely related to an onset response at around 100 ms poststimulus. The phonetic features, however, also show a second broader peak at around 400 ms, which is not present for the spectrogram representation.

To visualize the scalp regions underlying these components, we calculated the correlation coefficient between each component from the AV models for each time lag with each scalp electrode of the AV EEG. This gives a sense of how strongly the data on each channel has contributed to that component. The spatial pattern for the first component revealed strong contributions from channels over central and temporal scalp for speech in quiet for both the spectrogram and phonetic feature representations. For speech in noise there was an additional correlation with occipital channels, possibly indicating an increased contribution from visual areas to multisensory speech processing in noisy conditions. Occipital channels also made clear contributions for the second and third components for both conditions, however because of the lack of a clear temporal response for these components, we are hesitant to overinterpret this.

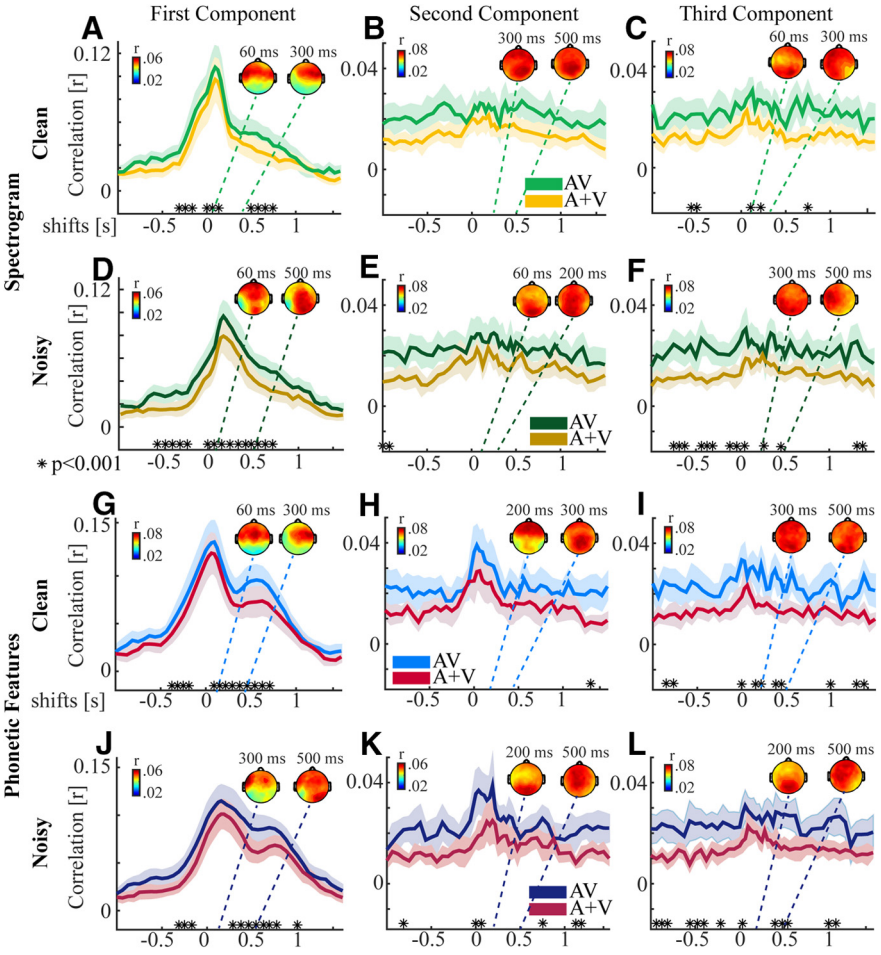

Figure 3. Correlations for the first three CCs using the spectrogram and phonetic feature representation of the speech stimulus at single time shifts. ***A–C***, The correlations using the spectrogram representation for the first three components using the AV model and the A + V model for speech in quiet, and ***D–F*** for speech in noise. ***G–I***, The single time shift correlations for the phonetic feature representation between the first three components for the AV and A + V model with the original raw AV EEG data for speech in quiet, and ***J–L*** for speech in noise. The respective topographies inset show the corresponding spatial correlations between the corresponding component of the AV model and the AV EEG; $*p < 0.001$ FDR corrected. The shaded region of the line plot marks the 95% confidence interval around the mean.

Isolating multisensory effects at the spectrotemporal and phonetic levels

As discussed above, the spectrogram and phonetic feature representations are highly correlated with each other. As such, measures of how well each individual representation maps to the EEG (as in Fig. 2) are difficult to interpret in terms of multisensory effects at specific hierarchical levels. To pinpoint effects at each specific level, we need to identify the unique contributions of the spectrogram and the phonetic feature representations to the EEG data. To do this, we first used the forward TRF model (Crosse et al., 2016a) to predict the EEG using one stimulus representation (e.g., the spectrogram). Time lags of 0–500 ms were used in the TRF analysis to match the delays used in the CCA analysis and to optimize the TRF model performance, we conducted a parameter search (over the range $\lambda = 2^{-2}, 2^{-1}, \ldots 2^{29}, 2^{30}$) for the λ value that maximized the correlation between the original and predicted EEG for each trial. This was done to ensure that we used the best possible prediction of the EEG in the partialling out procedure because we wanted to remove as much information as possible about the representation that was being partialled out. Then we subtracted this predicted EEG from the original EEG signal. Then we fed this residual EEG into the CCA analysis previously described. We performed this analysis for all stimulus representations. To ensure that there were no responses related to the stimulus feature which was regressed out using the TRF which could be extracted by CCA, we re-ran the CCA using the spectrogram and EEG with the spectrogram regressed out. We performed a similar analysis for the phonetic features. In both cases, we found that the correlations are extremely small and close to zero, which so leads us to believe that the partialling out was effective (result not shown; Extended Data Fig. 4-1).

This should isolate the unique contribution (if any) provided by the phonetic feature representation. Examining such a measure across our two conditions (clean speech and noisy speech) allowed us to test the hypothesis that multisensory integration effects should be particularly pronounced at the phonetic feature level for speech in noise. We also performed a similar analysis for the spectrogram representation to test its unique contribution to the multisensory effect in quiet and noise. In this case, we regressed out the phonetic feature representation from the EEG and then related the residual EEG to the spectrogram using CCA. Again, we did this for both speech in quiet and speech in noise, to test for interaction effects on our multisensory integration measures between acoustic and articulatory representations and speech in quiet and noise.

We limited our analysis here to the first CC because of its dominant role in the above results, as well as to the fact that it displays a greater consistency across subjects relative to the other components (see below, Consistency of CCs across subjects).

We found that multisensory gain at the level of acoustic processing (unique contribution from spectrogram) was significantly greater than zero for both clean speech and noisy speech (Fig. 4*A*). However, there was no difference in this measure between conditions (Fig. 4*B*; $p = 0.75$), and so we cannot state whether multisensory integration at the earliest cortical stages differs for speech in quiet and noise. Meanwhile, multisensory gain at the level of articulatory processing (unique contribution from phonetic features) was also significantly greater than zero for both clean speech and speech in noise (Fig. 4*C*). Importantly however, in line with our original hypothesis, there was a significant difference in this measure between conditions, with MSI gain being larger for speech in noise than speech in quiet (Fig. 4*D*; $p = 0.04$). This supports the idea that, when speech is noisy, the impact of complementary visual articulatory information on phonetic feature encoding is enhanced.

We used R (R Core Team, 2016) and lme4 (Bates et al., 2014) to perform a linear mixed effects analysis (Winter, 2013) with fixed effects of model type (AV vs A + V), stimulus representation (spectrogram or phonetic features) and environment (quiet or noisy) and random effect of subjects. In summary, we found a main effect of model type (driven by larger component correlations for AV vs A + V, $p < 0.0001$), and an interaction between stimulus representation and environment (driven by a decrease in correlation values for speech in noise vs speech in quiet for phonetic features, whereas there is no such decrease in correlations for the spectrogram representation across speech conditions, $p < 0.0001$); however, the three-way interaction was not significant ($p = 0.72$).

We also related the target word detection performance to the multisensory gains for both representations. To do this, we used F1 scores, which are calculated as the harmonic mean of precision and recall. This allowed us to investigate whether the probability of detecting target words in the multisensory condition exceeded the statistical facilitation produced by the unisensory stimuli (Stevenson et al., 2014). For more information on how this was calculated, see Crosse et al. (2016b). However, for both representations there was no correlation across subjects between the behavioural gain in target word detection and the multisensory gain calculated from the EEG using CCA (spectrogram: $r = 0.004$, $p = 0.98$; phonetic features: $r = 0.04$, $p = 0.86$).

It is difficult to isolate phoneme specific responses from the purely acoustic driven features, given their tightly linked association. To address the possibility that other forms of acoustic representations could explain this result, we re-ran two separate analyses using the half-wave rectified spectrogram derivative (Daube et al., 2019) and phonetic onsets (Brodbeck et al., 2018) in place of the phonetic features to test whether these representations would lead to similar results as for the phonetic features.

Using the spectrogram derivative, we found no difference in the gain between quiet and noisy speech conditions (Fig. 4*F*; $p = 0.1$). Similarly, in the case of phonetic onsets we found no effect (Fig. 4*H*; $p = 0.4$).

We also related the phonetic features to EEG that had both the spectrogram and spectrogram derivative partialled out. In this case, we found that there was no longer a significant difference in the gain between quiet and noisy speech conditions (Fig. 4*J*; $p = 0.25$). This is likely because of the fact that the size of the original effect is small because of it coming from differences between two models that are expected to be very similar in the first place, i.e., the AV and A + V models. Therefore regressing out these other representations reduces the correlation values further, resulting in a reduction in sensitivity to small effects. On top of this, we are comparing across different subjects, making our statistics less sensitive than would be the case for a within-subject design. Nevertheless, from Figure 4*C*, it is clear that the phonetic feature representation has noticeably higher correlation values compared with the correlation values for the other representations. This shows that phonetic features are explaining more variance in the EEG.

We then wanted to examine whether the multisensory integration effects at the phonetic feature level might be driven by specific phonemes. More precisely, we wondered whether the effect might be primarily driven by phonemes whose accompanying visual articulations are particularly informative (e.g., /b/, /p/, or /f/ compared with /g/ or /k/). To do this, we tested which

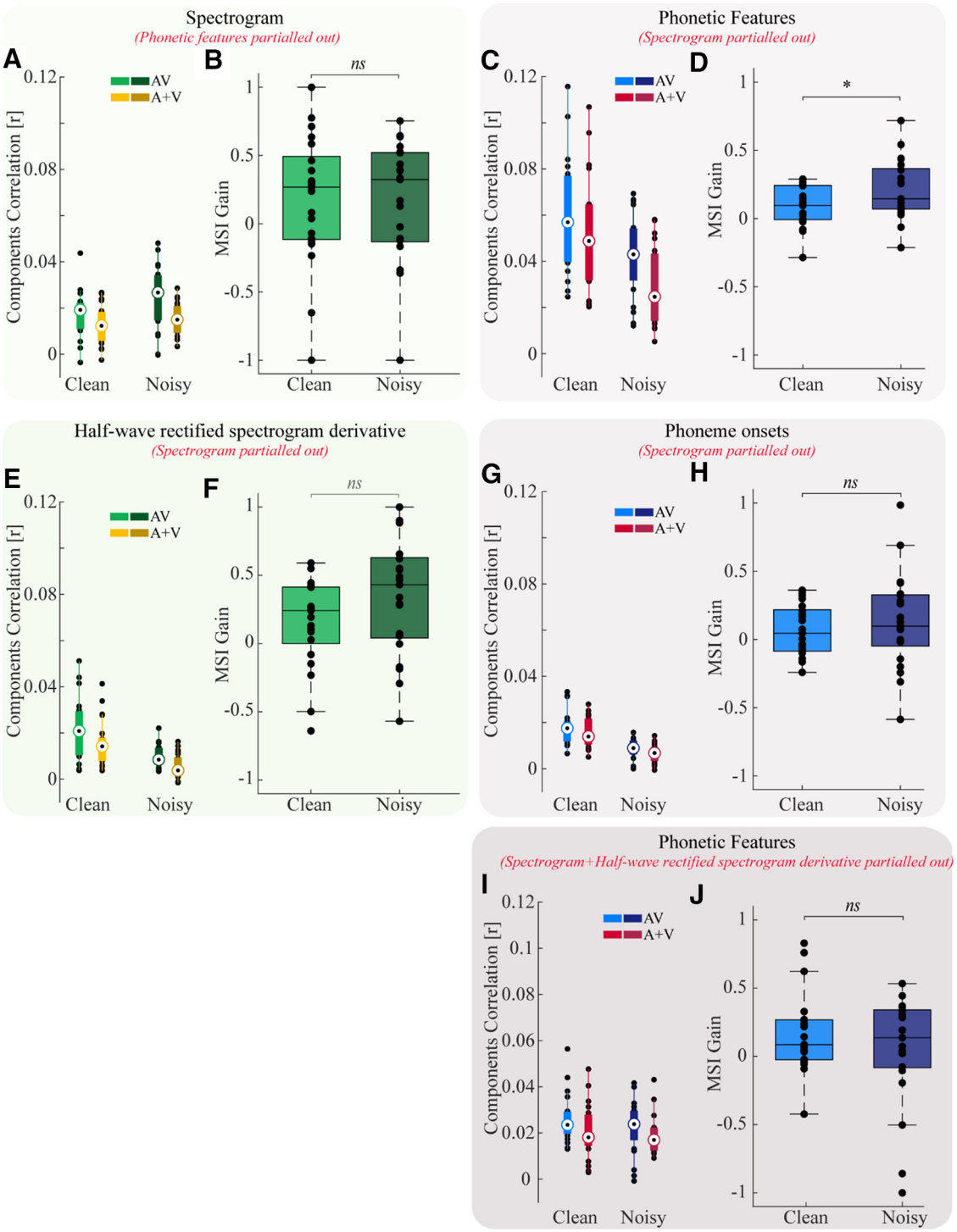

Figure 4. Multisensory gain for different speech representations for speech in quiet and in noise. ***A***, ***B***, For the unique spectrogram representation (phonetic features partialled out) there is no difference in gain, $p = 0.75$. ***C***, ***D***, For the unique phonetic features (spectrogram partialled out), we find a difference in gain between conditions, $p = 0.04$. ***E***, ***F***, Extended Data Figure 4-1 shows the effectiveness of the partialling out procedure, please refer to this figure for more information. Using the half-wave rectified spectrogram derivative we found no difference in gain between quiet and noisy conditions ($p = 0.1$) and similarly for phonetic onsets there was no difference across conditions (***G***, ***H***; $p = 0.4$). Using the phonetic features after partialling out the spectrogram and spectrogram derivative, there is no difference in gain across quiet and noisy speech conditions (***I***, ***J***; $p = 0.25$). Two-tailed unpaired permutation tests were used throughout. The boxplot function with compact format from MATLAB was used for figure parts ***A***, ***C***, ***E***, ***G***. The edges of the boxplots display the 25th and 75th percentiles and the median is marked by a black dot inside a white circle. The whiskers extend to the most extreme data points that the algorithm does not consider to be outliers and the outliers are plotted individually. For figure parts ***B***, ***D***, ***F***, ***H*** the default boxplot from MATLAB is used. The only difference in this case is that the median is displayed using a black horizontal line.

phonemes were most correlated with the first CC. This involved taking the first component arising from the rotation of the phonetic feature stimulus on the left-out data and then calculating the correlation between this component and the time series of each phoneme. If there is a high correlation between the component and a particular phoneme then it suggests that this phoneme is strongly represented in that component and it plays a role in driving the stimulus-EEG correlations that we have

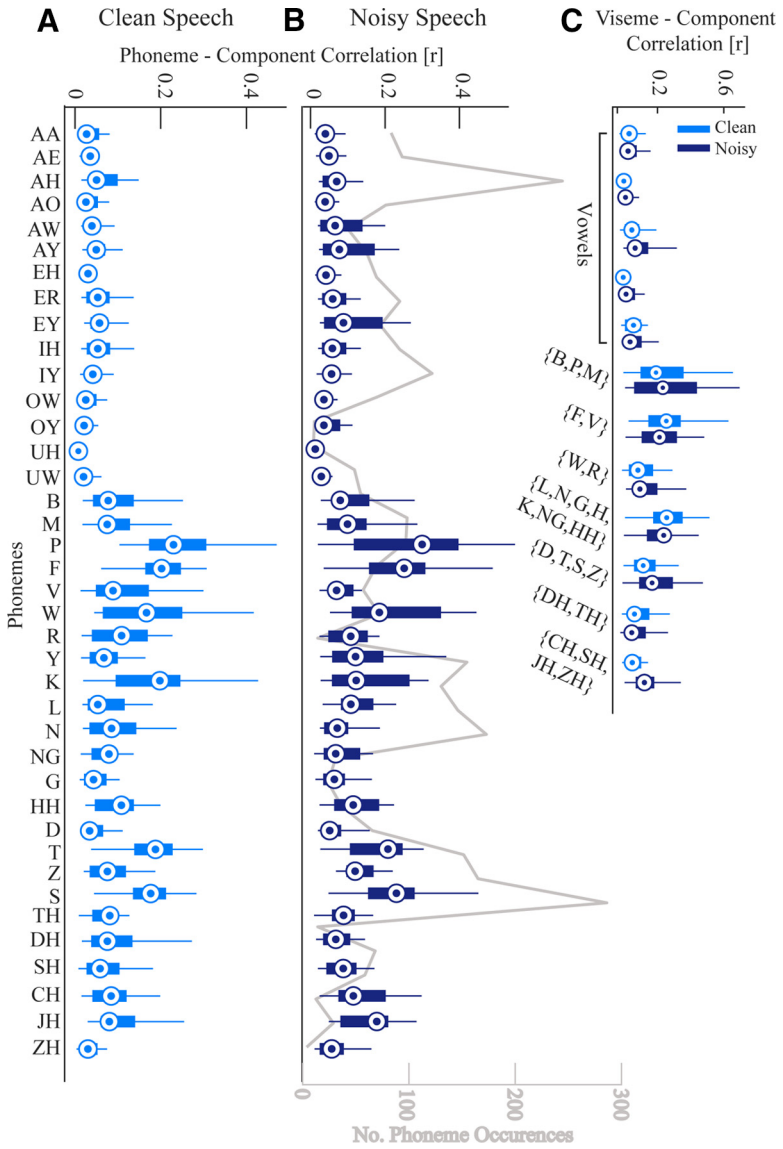

Figure 5. Correlation between phonemes and visemes time series with the first CC. ***A***, The correlations for each phoneme from the clean speech data. ***B***, The correlations for each phoneme from the speech in noise data. The gray line plots the number of phoneme occurrences to show that it is not the case that the most frequent phonemes dominate the data. ***C***, The correlation between each viseme (groups of visually similar phonemes) and the first component. The compact version of MATLAB's boxplot function is used here (for description, see Fig. 4 caption).

reported here. This analysis revealed that phonemes such as /p/, /f/, /w/, and /s/were most strongly represented in the first component. In general, it also showed that consonants were more correlated with the component than vowels (Fig. 5*A*,*B*), although for speech in noise this effect was slightly less pronounced (Fig. 5*B*). To see these results in terms of visual articulatory features, we grouped the phonemes into visemes (the visual analog of phonemes), based on the mapping defined previously (Auer and Bernstein, 1997). This showed that bilabials (/b/, /p/, and /m/) and labio-dentals (/f/ and /v/) were the features most correlated with the first component. Finally, we also checked that these phoneme-component correlations were not simply explainable as a function of the number of occurrences of each phoneme. To check this, we tested for a relationship between the phoneme-component correlations and the number of occurrences of each phoneme (Fig. 5*B*, gray line). No correlation between the two was found for either speech in quiet ($p = 0.99$) or speech in noise ($p = 0.71$).

This analysis highlights how different visual-phonetic features contribute to our multisensory effects for phonetic features. In particular we find that bilabials and labiodentals have the highest correlation with the first CC, suggesting that these features contribute most to the effects shown here. This is line with early work examining the saliency of visual phonetic features which found that place of articulation (i.e., the ability to distinguish between labials, e.g., /p/, /b/, and non-labials, e.g., /d/, /t/) was the most salient visual phonetic feature, followed by manner of articulation (i.e., distinguishing between stops, e.g., /d/ and fricatives, e.g., /f/) and voicing (Walden et al., 1977).

Consistency of CCs across subjects

CCA finds stimulus-EEG matrix rotations on a single subject basis. As such, for us to make general conclusions about results gleaned from individual CCs, we must examine how similar the individual components are across subjects. To do this, we took the components for each subject and calculated the correlation (using Pearson's *r*) for every subject pair (Fig. 6). Given its dominant role in capturing EEG responses to speech, and in the results we have presented above, we were particularly interested in consistency of component one across subjects.

For the spectrogram, the first components of the AV and A + V models were significantly more correlated across subjects than all other components for clean speech ($p < 0.0001$ for both). For speech in noise the first component of the AV model was not significantly more correlated across subjects than the second component ($p = 0.055$) but it had a significantly higher correlation than the remainder of the components ($p < 0.0001$). The first A + V component for speech in noise was significantly more correlated across subjects than all others ($p < 0.001$).

For the phonetic features model, a similar pattern emerged for the clean speech condition with the first components of the AV and A + V models being significantly more correlated across subjects than all other components ($p < 0.0001$ for both). For noisy speech, the first component of the AV model was again significantly better than all others ($p < 0.02$) and similarly the first component of the A + V model had a higher correlation than all others ($p < 0.0001$).

Altogether, we found that the first component is most correlated across subjects, while later components are less correlated. This suggests that the first component, which dominated our results above, is also the most consistent component across subjects and, thus, that it is capturing processes that are general across those subjects. In contrast, the later components, as well as being smaller, are more variable across subjects, and, accordingly, may not be capturing similar underlying processes. This in turn can make results from these later components more difficult to interpret.

Discussion

In this work, we have used a CCA-based framework for relating multivariate stimulus representations to multivariate neural data

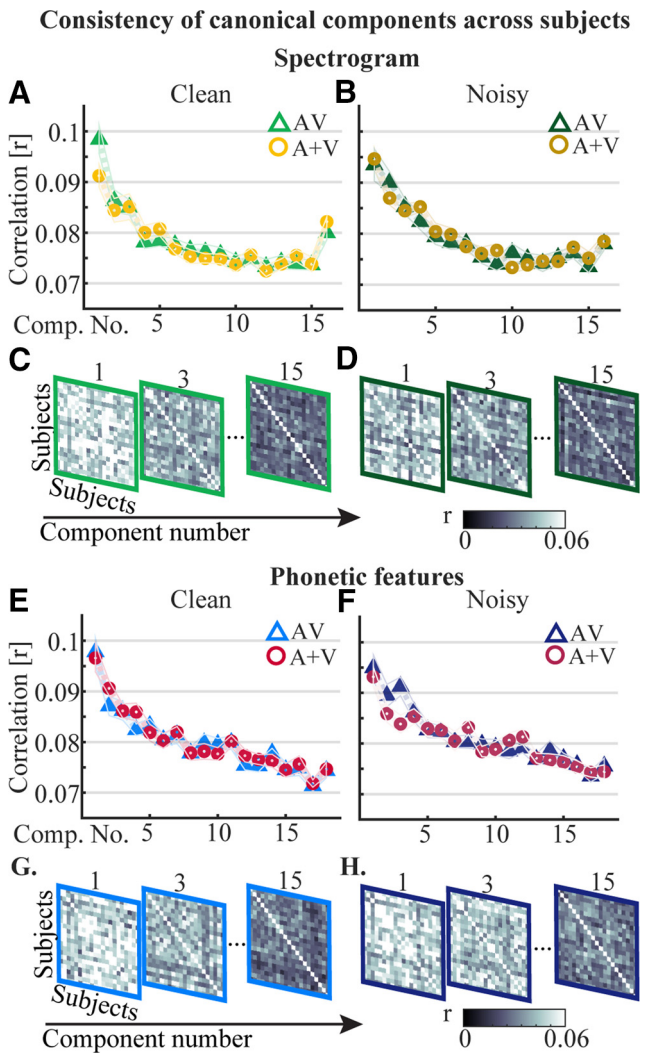

Figure 6. Consistency of CCs across subjects for the spectrogram and the phonetic features after partialling out the other representation. ***A***, ***B***, Correlations of AV and A + V CCs across subjects for the spectrogram representation for speech in quiet and speech in noise, respectively. ***C***, ***D***, Correlation matrices visualizing the reduction in consistency in the AV component activity across subjects as the component number increases for the spectrogram. ***E***, ***F***, The correlation of the CCs for the phonetic feature representation across subjects. ***G***, ***H***, Corresponding correlation matrices for speech in quiet and speech in noise, respectively.

to study the neurophysiological encoding of multidimensional acoustic and linguistic features of speech. Our results show significant AV integration effects on the encoding of the spectrogram and phonetic features of both clean and noisy speech. This provides evidence for the idea that AV speech integration is a multistage process and that vision interacts with speech processing at multiple stages.

Enhanced multisensory integration effects at the phonetic level of processing for speech in noise

It is well known that the enhancement of auditory speech processing provided by visual speech varies with listening conditions (Ross et al., 2007). However, the details of how visual speech impacts auditory speech processing at different hierarchical levels remains to be fully elucidated. There is a growing body of evidence indicating that AV speech integration likely occurs over multiple stages (Peelle and Sommers, 2015). In particular, it is thought that visual speech provides temporal information about the acoustic speech which can affect the sensitivity of auditory cortex (Grant and Seitz, 2000; Okada et al., 2013), as well as provide complementary cues that contain articulatory information which may be integrated with acoustic information in STS (Beauchamp et al., 2004; Kayser and Logothetis, 2009; Nath and Beauchamp, 2011).

In this study, we found evidence that AV speech integration may operate differently under different listening conditions [clean vs noisy (−9 dB) speech]. Specifically, for the encoding of low-level spectrogram features, we found that the integration effects are substantial for both speech in quiet and speech in noise. These integration effects are likely to be primarily driven by modulations of responses in early auditory cortex by temporal information provided by visual speech, which often precedes the auditory speech (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014). This result is also in line with recent work demonstrating multisensory benefits at the spectrotemporal level elicited by a visual stimulus that did not contain articulatory detail, dissociating the effect from access to higher-level articulatory details (Plass et al., 2019). Furthermore, the lack of any difference in the magnitude of these integration effects between clean and noisy speech conditions suggests that the benefits of visual speech provided at a low level of processing might be similar regardless of acoustic conditions.

Using a higher-level phonetic feature representation, we found that the AV integration effects vary weakly depending on the acoustic conditions (after regressing out the contribution of the spectrogram). Specifically, we found larger integration effects for phonetic feature encoding in noisy speech than in clean speech. We suggest that this benefit is likely to be driven by an increased reliance on the visual articulations which help the listener to understand the noisy speech content by constraining phoneme identity (Karas et al., 2019). In line with this, we also show that the phonemes that most contribute to these results are those that have particularly informative visual articulations (Fig. 5). Nevertheless, this effect is weakened by the removal of additional speech features (i.e., the spectrogram and half-wave spectrogram derivative). Future work will be required using a within-subject design to investigate how integration at different stages may or may not vary with environment conditions.

While recent research has challenged the notion that scalp recorded responses to speech reflect processing at the level of phonemes (Daube et al., 2019), our findings reveal a dissociation in AV integration effects on isolated measures of acoustic and phonetic processing across listening conditions. This seems difficult to explain based on considering acoustic features alone and seems consistent with the idea of visual articulations influencing the categorization of phonemes (Holt and Lotto, 2010). More generally, we take this as a further contribution to a growing body of evidence for phonological representations in cortical recordings to naturalistic speech (Di Liberto et al., 2015; Khalighinejad et al., 2017; Brodbeck et al., 2018; Yi et al., 2019; Gwilliams et al., 2020).

One brain region likely involved in exploiting the articulatory information when the speech signal is noisy is STS, which has been shown to have increased connectivity with visual cortex in noisy compared with quiet acoustic conditions (Nath and Beauchamp, 2011). While it remains an open question as to how much speech-specific processing is performed by visual cortex (Bernstein and Liebenthal, 2014), there is some early evidence supporting the notion that visual cortex might processes speech at the level of categorical linguistic (i.e., phonological) units (O'Sullivan et al., 2017; Hauswald et al., 2018). If true, visual cortex would be in a position to relay such categorical, linguistic

information to directly constrain phoneme identity, again, possibly in STS. On top of this it has been shown that frontal cortex selectively enhances processing of the lips during silent speech compared with when the auditory speech is present, suggesting an important role for visual cortex in extracting articulatory information from visual speech cues (Ozker et al., 2018). Thus, it is plausible that the greater multisensory gain seen here for phonetic features when the speech is noisy is underpinned by an enhancement of mouth processing in visual cortex which feeds information about the articulations to STS where they influence the online processing of the acoustic speech. However, follow-up studies are required to conclusively demonstrate such an effect occurring.

Investigating hierarchical stages of speech processing, CCA captures relationships between multidimensional stimuli and EEG

The ERP technique has for a long time been used to advance our understanding of the multisensory integration of speech (Stein and Meredith, 1993; Molholm et al., 2002; Klucharev et al., 2003; van Wassenhove et al., 2005; Saint-Amour et al., 2007; Bernstein et al., 2008; Shahin et al., 2018). However, this approach is ill suited for use with natural, continuous speech stimuli.

More recently, researchers have begun to use methods such as multivariate regression (Crosse et al., 2016a) in the forward direction (predicting neural data from the stimulus; Lalor and Foxe, 2010; Ding and Simon, 2012; Zion Golumbic et al., 2013; Di Liberto et al., 2015; O'Sullivan et al., 2017; Broderick et al., 2018) and backward direction (stimulus reconstruction from neural data; Mesgarani et al., 2009; Crosse et al., 2015a,b, 2016b), which allows characterization of neural responses to natural speech. However, regression models in their general form allow only univariate-multivariate comparison, whereas with CCA one can relate multivariate stimulus representations (discrete/continuous) to multivariate neural responses. This is a useful advance over current techniques to study speech processing since CCA can use all features (of the stimulus and the neural response data) simultaneously to maximize the correlation between the speech representation and the neural data (de Cheveigné et al., 2018). Importantly, this approach has allowed us to answer questions which we could not do with previous methods, such as the impact of visual speech on auditory speech processing at different stages.

Our results show significant multisensory interaction effects in EEG responses based on the spectrogram and phonetic feature representations of the speech signal and so provides support for the multistage framework for AV speech integration. Examining the relationship between the stimulus representations and EEG data at individual time shifts reveals a peak in the correlation at around 100 ms poststimulus for both the spectrogram and phonetic feature representations. This is likely attributable to a sound onset response. For the phonetic feature representation however, there is also a second broad peak at around 300–600 ms, whereas for the spectrogram, there is no noticeable second peak. In terms of the scalp regions which most contribute to the first component, we found it to be dominated by central and temporal regions for speech in quiet, and for speech in noise there is a greater contribution from more parietal and occipital regions. This is likely because of increased contributions from the visual areas when the acoustic speech is noisy.

Limitations and future considerations

The use of CCA to study responses to natural speech has allowed us to answer questions that we could not previously answer. The use of natural and continuous stimuli is important to study the neural systems involved in processing AV speech in the real world (Hamilton and Huth, 2018). However, there are some drawbacks in the experiment design which could be improved on in the future. The current paradigm is made to be somewhat unnatural by the presentation of A, V, and AV speech, each in separate 1-min trials. It is possible therefore, that in the V condition, subjects find it very difficult to understand the speech and so may result in a decrease in attention to the speech material (or an increase for those subjects that are trying harder). If attention to the speech in V trials differs in comparison with attention to the visual aspect of the AV speech stimulus, then this could impact our V EEG responses that are added to the auditory-only EEG responses to generate an A + V EEG response. This issue is common to almost all studies that examine multisensory integration effects of AV speech (and indeed many multisensory experiments more generally).

One approach that has recently been suggested to overcome this is the use of AV speech in every trial but varying the delay between the auditory and visual speech across trials such that the AV speech is still intelligible. This variability in delay can theoretically allow one to characterize the unisensory responses using deconvolution (Metzger et al., 2020). This approach was used for the presentation of single words but would be interesting to apply it in a paradigm using continuous speech.

Nevertheless, the effect of interest in this study is the relative change in the difference between the performance of an AV and A + V model applied to AV EEG responses across quiet and noisy speech conditions. We would not expect the difference in performance to change with different speech conditions if the improvement of the AV model was simply driven by a poor A + V model because of a poor contribution from the V-only response in the A + V EEG response.

In conclusion, this work has used a novel framework to study multisensory interactions at the acoustic and phonetic levels of speech processing. This has revealed that multisensory effects are present for both the spectrogram and phonetic feature representations when the speech is in quiet or when it is masked by noise. There is also evidence to suggest that these multisensory interactions may vary with listening conditions, however, future work will be required to examine this question in more detail.

References

Auer ET Jr, Bernstein LE (1997) Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J Acoust Soc Am 102:3704–3710.

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.

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. arXiv 1406.5823.

Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.

Berman AL (1961) Interaction of cortical responses to somatic and auditory stimuli in anterior ectosylvian gyrus of cat. J Neurophysiol 24:608–620.

Bernstein LE, Liebenthal E (2014) Neural pathways for visual speech perception. Front Neurosci 8:386.

Bernstein LE, Auer ET Jr, Wagner M, Ponton CW (2008) Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39:423–435.

Besle J, Fischer C, Bidet-Caulet A, Lecaignard F, Bertrand O, Giard M-H (2008) Visual activation and audiovisual interactions in the auditory

cortex during speech perception: intracranial recordings in humans. J Neurosci 28:14301–14310.
Bilenko NY, Gallant JL (2016) Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. Front Neuroinform 10:49.
Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.e5.
Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803–809.e3.
Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596.
Campbell R (2008) The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc Lond B Biol Sci 363:1001–1010.
Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5: e1000436.
Chomsky N, Halle M (1968) The sound pattern of English. New York: Harper and Row.
Combrisson E, Jerbi K (2015) Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 250:126–136.
Crosse MJ, Butler JS, Lalor EC (2015a) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J Neurosci 35:14195–14204.
Crosse MJ, ElShafei HA, Foxe JJ, Lalor EC (2015b) Investigating the temporal dynamics of auditory cortical activation to silent lipreading. 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), pp 308–311. Montpellier, France. IEEE.
Crosse MJ, Di LG, Bednar A, Lalor EC (2016a) The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604.
Crosse MJ, Di Liberto GM, Lalor EC (2016b) Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J Neurosci 36:9888–9895.
Cruz-Cano R, Lee MLT (2014) Fast regularized canonical correlation analysis. Comput Stat Data Anal 70:88–100.
Daube C, Ince RAA, Gross J (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr Biol 29:1924–1937.e9.
de Cheveigné A, Arzounian D (2018) Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. Neuroimage 172:903–912.
de Cheveigné A, Wong DDE, Di Liberto GM, Hjortkjær J, Slaney M, Lalor E (2018) Decoding the auditory brain with canonical component analysis. Neuroimage 172:206–216.
Delorme A, 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.
DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci USA 109:E505–E514.
Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.
Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89.
Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735.
Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88:41–46.
Eskelund K, Tuomainen J, Andersen TS (2011) Multistage audiovisual integration of speech: dissociating identification and detection. Exp Brain Res 208:447–457.
Gorman K, Howell J, Wagner M (2011) Prosodylab-Aligner: a tool for forced alignment of laboratory speech. Can Acoust 39:2.
Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197–1208.
Gwilliams L, King JR, Marantz A (2020) Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. bioRxiv 2020.2004.2004.025684.
Hamilton LS, Huth AG (2018) The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci 35:573–510.
Hauswald A, Lithari C, Collignon O, Leonardelli E, Weisz N (2018) A visual cortical network for deriving phonological information from intelligible lip movements. Curr Biol 28:1453–1459.e3.
Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.
Holt LL, Lotto AJ (2010) Speech perception as categorization. Atten Percept Psychophys 72:1218–1227.
Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377.
Irino T, Patterson RD (2006) A dynamic compressive gammachirp auditory filterbank. IEEE Trans Audio Speech Lang Process 14:2222–2232.
Karas PJ, Magnotti JF, Metzger BA, Zhu LL, Smith KB, Yoshor D, Beauchamp MS (2019) The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife 8:e48116.
Kayser C, Logothetis NK (2009) Directed interactions between auditory and superior temporal cortices and their role in sensory integration. Front Integr Neurosci 3:7.
Khalighinejad B, Cruzatto da Silva G, Mesgarani N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. J Neurosci 37:2176–2185.
Klucharev V, Möttönen R, Sams M (2003) Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Res Cogn Brain Res 18:65–75.
Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31:189–193.
Leurgans SE, Moyeed RA, Silverman BW (1993) Canonical correlation analysis when the data are curves. J R Stat Soc Series B Stat Methodol 55:725–740.
Luo H, Liu Z, Poeppel D (2010) Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol 8:e1000445.
Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102:3329–3339.
Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010.
Metzger BA, Magnotti JF, Wang Z, Nesbitt E, Karas PJ, Yoshor D, Beauchamp MS (2020) Responses to visual speech in human posterior superior temporal gyrus examined with iEEG deconvolution. J Neurosci 40:6938–6948.
Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 14:115–128.
Nath AR, and Beauchamp MS (2011) Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. J Neurosci 31:1704–1714.
O'Sullivan AE, Crosse MJ, Di Liberto GM, Lalor EC (2017) Visual cortical entrainment to motion and categorical speech features during silent lipreading. Front Hum Neurosci 10:679.
Okada K, Venezia JH, Matchin W, Saberi K, Hickok G (2013) An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PLoS One 8:e68959.
Ozker M, Yoshor D, Beauchamp MS (2018) Frontal cortex selects representations of the talker's mouth to aid in speech perception. Elife 7:e30387.
Parsons TW (1987) Voice and speech processing. New York: McGraw-Hill College.
Peelle JE, Sommers MS (2015) Prediction and constraint in audiovisual speech perception. Cortex 68:169–181.
Plass J, Brang D, Suzuki S, Grabowecky S (2019) Vision perceptually restores auditory spectral dynamics in speech. Proc Natl Acad Sci USA 117:16920–16927.
R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147–1153.

Saint-Amour D, De Sanctis P, Molholm S, Ritter W, Foxe JJ (2007) Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia 45:587–597.

Schwartz JL, 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 Comput Biol 10:e1003743.

Schwartz JL, Berthommier F, and Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93:B69–B78.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107.

Shahin AJ, Backer KC, Rosenblum LD, Kerlin JR (2018) Neural mechanisms underlying cross-modal phonetic encoding. J Neurosci 38:1835–1849.

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge: The MIT Press.

Stevenson RA, Ghose D, Fister JK, Sarko DK, Altieri NA, Nidiffer AR, Kurela LR, Siemann JK, James TW, Wallace MT (2014) Identifying and quantifying multisensory integration: a tutorial review. Brain Topogr 27:707–730.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331.

Tye-Murray N, Spehar B, Myerson J, Sommers MS, Hale S (2011) Cross-modal enhancement of speech detection in young and older adults: does signal content matter? Ear Hear 32:650–655.

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 102:1181–1186.

Vinod HD (1976) Canonical ridge and econometrics of joint production. J Econom 4:147–166.

Walden BE, Prosek RA, Montgomery AA, Scherr CK, Jones CJ (1977) Effects of training on the visual recognition of consonants. J Speech Hear Res 20:130–145.

Winter B (2013) Linear models and linear mixed effects models in R: tutorial 11. arXiv 1308.5499.

Yi HG, Leonard MK, Chang EF (2019) The encoding of speech sounds in the superior temporal gyrus. Neuron 102:1096–1110.

Yuan J, Liberman M (2008) Speaker identification on the SCOTUS corpus. J Acoust Soc Am 123:3878.

Zhu LL, Beauchamp MS (2017) Mouth and voice: a relationship between visual and auditory preference in the human superior temporal sulcus. J Neurosci 37:2697–2708.

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D (2013) Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J Neurosci 33:1417–1426.