Behavioral/Cognitive

Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment

Emily S. Teoh,[1] Farhin Ahmed,[2] and Edmund C. Lalor[1,2]

[1]School of Engineering, Trinity Centre for Biomedical Engineering, and Trinity College Institute of Neuroscience, Trinity College, University of Dublin, Dublin 2, Ireland, and [2]Department of Neuroscience, Department of Biomedical Engineering, and Del Monte Neuroscience Institute, University of Rochester, Rochester, New York 14627

Humans have the remarkable ability to selectively focus on a single talker in the midst of other competing talkers. The neural mechanisms that underlie this phenomenon remain incompletely understood. In particular, there has been longstanding debate over whether attention operates at an early or late stage in the speech processing hierarchy. One way to better understand this is to examine how attention might differentially affect neurophysiological indices of hierarchical acoustic and linguistic speech representations. In this study, we do this by using encoding models to identify neural correlates of speech processing at various levels of representation. Specifically, we recorded EEG from fourteen human subjects (nine female and five male) during a "cocktail party" attention experiment. Model comparisons based on these data revealed phonetic feature processing for attended, but not unattended speech. Furthermore, we show that attention specifically enhances isolated indices of phonetic feature processing, but that such attention effects are not apparent for isolated measures of acoustic processing. These results provide new insights into the effects of attention on different prelexical representations of speech, insights that complement recent anatomic accounts of the hierarchical encoding of attended speech. Furthermore, our findings support the notion that, for attended speech, phonetic features are processed as a distinct stage, separate from the processing of the speech acoustics.

Key words: attention; cocktail party; EEG; speech

Significance Statement

Humans are very good at paying attention to one speaker in an environment with multiple speakers. However, the details of how attended and unattended speech are processed differently by the brain is not completely clear. Here, we explore how attention affects the processing of the acoustic sounds of speech as well as the mapping of those sounds onto categorical phonetic features. We find evidence of categorical phonetic feature processing for attended, but not unattended speech. Furthermore, we find evidence that categorical phonetic feature processing is enhanced by attention, but acoustic processing is not. These findings add an important new layer in our understanding of how the human brain solves the cocktail party problem.

Introduction

The ability to focus on a single talker amid multiple sounds is essential for human communication. Cherry (1953) brought this phenomenon to the fore, investigating our behavioral capacity for selective auditory attention. His study stimulated subsequent inquiries and the proposal of various theories of attention (Broadbent, 1958; Moray, 1959; Deutsch and Deutsch, 1963; Treisman, 1964; Johnston and Wilson, 1980). These theories model attention as a selective filter that rejects the unattended message beyond a certain information processing stage. However, whether this filter operates at an early or late stage of speech processing is still unresolved.

In recent years, neuroscience has sought to contribute to this debate by leveraging our increased understanding of the hierarchical nature of speech processing in the human cortex. In

Received June 8, 2020; revised Sep. 28, 2021; accepted Sep. 29, 2021.

Author contributions: E.S.T. and E.C.L. designed research; E.S.T. and F.A. performed research; E.S.T. and F.A. analyzed data; E.S.T. wrote the first draft of the paper; F.A. and E.C.L. edited the paper; F.A. and E.C.L. wrote the paper.

This work was supported by the Science Foundation Ireland Career Development Award 15/CDA/3316, the Irish Research Council Government of Ireland Postgraduate Scholarship GOIPG/2015/3378, the Del Monte Neuroscience Institute at the University of Rochester, and the National Institutes of Health Grant 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.1455-20.2021

Copyright © 2022 Teoh et al.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

particular, to construe meaning from sound, the cortex is posited to compute several intermediate levels of increasingly abstract representations in different functionally specialized regions of a hierarchically organized cortical network (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Peelle et al., 2010; DeWitt and Rauschecker, 2012; Kell et al., 2018). For example, single talker neuroimaging studies have revealed that core auditory regions code low-level features which are combined in higher areas to yield more abstract neural codes (Binder et al., 2000; Davis and Johnsrude, 2003; Okada et al., 2010). And in terms of multitalker selective attention, studies have shown that primary auditory cortical responses represent all talkers regardless of attentional state (O'Sullivan et al., 2019), but in "higher" areas such as the superior temporal gyrus (STG), only the attended speaker is represented (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013; O'Sullivan et al., 2019). Additionally, EEG/MEG studies examining the latency of neural responses (which can be considered a rough proxy measure of processing at different hierarchical stages) have shown that all talkers are co-represented in early components, with distinct responses to the attended speaker appearing only in later components (Ding and Simon, 2012b; Power et al., 2012; Puvvada and Simon, 2017). Importantly, this specificity is not something that is necessarily true for more simplistic nonspeech stimuli and tasks (Power et al., 2011).

As well as understanding how attention affects processing in different brain areas and at different latencies, we also wish to understand how attention influences the encoding of the different speech representations that are thought to be computed by this hierarchical network. In recent years, encoding/decoding methods (Crosse et al., 2016; Holdgraf et al., 2017) have revealed neural indices of a number of these representations, from low-level acoustics to semantics, in brain responses to continuous natural speech (Ahissar et al., 2001; Mesgarani et al., 2014; Di Liberto et al., 2015, 2019; Tang et al., 2017; Brodbeck et al., 2018; Broderick et al., 2018; Daube et al., 2019; Teoh et al., 2019). And it has been shown that neural indices of lexical (Brodbeck et al., 2018) and semantic (Broderick et al., 2018) processing can only be found for attended speech, in contrast to acoustically-driven measures like those based on the amplitude envelope that are less affected by attention (Brodbeck et al., 2018). Taken together, the results from these anatomic, time-resolved, and representational approaches support the notion that higher-order regions (and higher-level representations) are more greatly modulated by attention.

In this study, we explore how attention modulates neural indices of speech processing at prelexical levels. Specifically, we aim to answer two main questions. Our first question is: does the brain process speech at the level of phonemes and, if so, does it do so for both attended and unattended speech? The notion of prelexical processing at the level of phonemes has received neuroscientific support in recent years (Chang et al., 2010; Mesgarani et al., 2014; Di Liberto et al., 2015; Khalighinejad et al., 2017; Brodbeck et al., 2018; Gwilliams et al., 2020) . But the idea has also been challenged, with some recent research suggesting that what appear to be categorical responses to phonetic features might actually be responses to simple acoustic features (Daube et al., 2019). Here, to address this debate, we explore brain responses to speakers with very different acoustics (i.e., male and female). The rationale here is that a greater increase in acoustic variability across the same phonemes will improve our ability to distinguish categorical brain responses to different phonetic features from brain responses driven by the speech acoustics. Our second question is: at what hierarchical levels does attention influence processing? We hypothesize that we will see stronger attention effects on isolated measures of phonetic feature processing than on isolated measures of acoustic processing. This would support the idea of hierarchical attention effects suggested by previous studies by showing a dissociation in the influence of attention on different aspects of sublexical processing.

Materials and Methods

Subjects

Fourteen subjects (nine female and five male) between the ages of 19 and 30 participated in the experiment. All subjects were right-handed and spoke English as their primary language. Subjects reported no history of hearing impairment or neurologic disorder. Each subject provided written informed consent before testing and received monetary reimbursement. The study was approved by the Research Subjects Review Board at the University of Rochester. Some of the data used here (the first 10 trials) were analyzed differently for a previous study (Teoh and Lalor, 2019).

Stimuli and procedures

Subjects undertook 40 1-min trials in two separate blocks. Stimuli consisted of two works of fiction, one narrated by a female talker and the other by a male talker. The stimuli were taken from two Sherlock Holmes novels: "A Study in Scarlet" (narrated by a female speaker) and "The Hound of the Baskervilles" (narrated by a male speaker). The stories were narrated by British speakers. Silent gaps in the audio exceeding 0.3 s were truncated to 0.3 s in duration.

The stimuli were filtered using head-related transfer functions (HRTFs), giving rise to the perception that the talkers were at 90° to the left and right of the subject. The HRTFs used were obtained from the CIPIC database (Algazi et al., 2001). Subjects were always instructed to attend to one of the two talkers, a counterbalanced paradigm was employed in which, over the course of the experiment, they would have attended to both male and female talkers and at both locations. In terms of content, the attended story segments were presented contiguously (i.e., each trial beginning where previous one ended), but the unattended story segments were presented in a random order. Each trial lasted exactly 60 s, with no special effort made to end the trial at a natural break in the story.

Before the experiment, subjects were asked to minimize motor activities and to maintain visual fixation on a crosshair centered on the screen during trials. After each trial, subjects were required to answer four multiple-choice comprehension questions on each of the stories (attended and unattended).

All stimuli were sampled with a frequency of 44,100 Hz and were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems.

Data acquisition and preprocessing

The experiment was conducted in a soundproof room. A Biosemi ActiveTwo system was used to record EEG data from 128 electrode positions on the scalp as well as two electrodes over the mastoid processes (all digitized at 512 Hz).

EEG data were re-referenced to the mastoids. Automatic bad channel rejection and interpolation was performed. A particular channel was deemed as bad if the standard deviation of the channel was lower than a third of or exceeded three times the mean of the standard deviation of all channels. In place of the bad channel, data were interpolated from the four nearest neighboring electrodes using spherical spline interpolation (Delorme and Makeig, 2004). To decrease subsequent processing time, data were downsampled to 128 Hz. Data were filtered between 0.2 and 8 Hz using a zero-phase shift FIR filter with Kaiser window (filter order was 20,000 for the 0.2-Hz high-pass filter and 1000 for the 8-Hz low-pass filter; maximum stopband attenuation was −60 dB). Low δ-band frequencies (down to 0.2 Hz) were included as they have previously been found to be important for speech processing (Teoh et al., 2019).

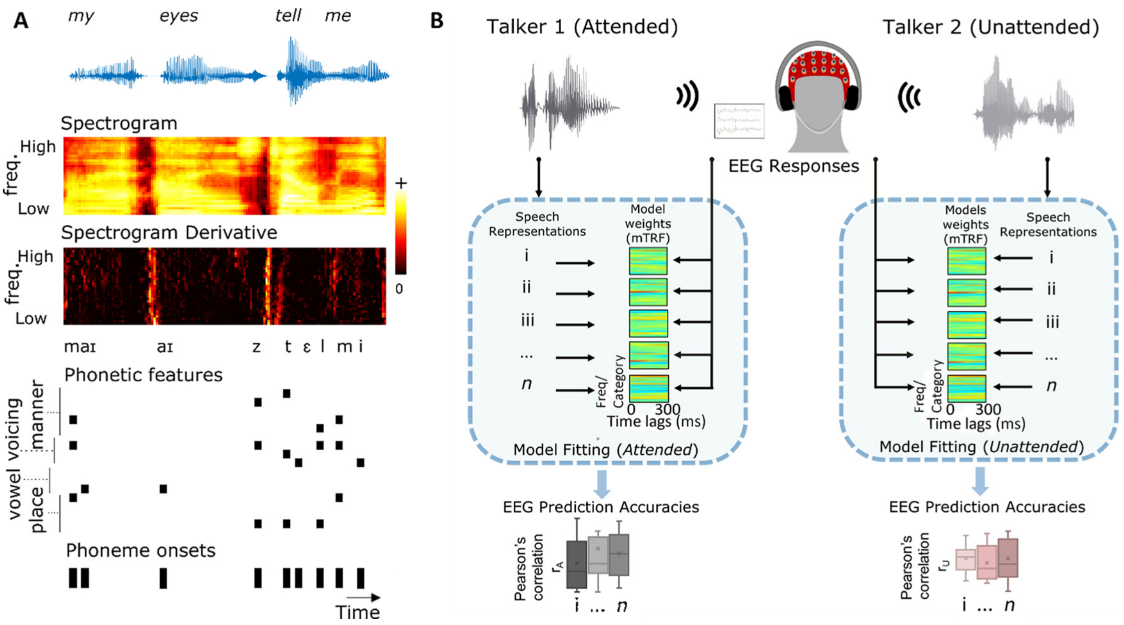

Figure 1. ***A***, Speech representations: the first row depicts an acoustic waveform of an excerpt taken from one of the stimuli. Subsequent rows show the computed acoustic and phonetic representations for that excerpt. ***B***, Analysis framework: cross-validation is used to train forward models mapping the different attended and unattended speech representations to EEG. These models are then used to predict left-out EEG. Pearson's correlation is used to evaluate model accuracy, and the prediction accuracies of acoustic and phonetic models are compared.

Speech representations

We computed acoustic and phonetic speech representations of both the attended and unattended stories (Fig. 1*A* depicts these representations for an excerpt taken from one of the audio clips).

Acoustic

- Spectrogram (*s*): the GBFB toolbox (Schädler and Kollmeier, 2015) was used to extract the spectral decomposition of the time-varying stimulus energy in 31 mel-spaced bands with logarithmic compressive nonlinearity. The spectrogram representation computed in this way is consistent with (Daube et al., 2019), where it was shown to better predict neural activity than other techniques.
- Spectrogram derivative (*sD*): the temporal derivative of each spectrogram channel was computed, and then half-wave rectified. This represents an approximation of the onsets in the acoustics and was shown to contribute to predicting neural responses in previous research (Daube et al., 2019).

Phonetic

- Phonetic features (*f*): phoneme-level segmentation of the stimuli was first computed using FAVE-Extract (Rosenfelder et al., 2014) and the Montreal Forced Aligner (McAuliffe et al., 2017) via the DARLA web interface (Reddy and Stanford, 2015). The phoneme representation was then mapped into a space of 19 features based on the mapping described previously (Chomsky and Halle, 1968; Mesgarani et al., 2014). The features describe the articulatory and acoustic properties of the phonetic content of speech (e.g., bilabial, plosive, voiced, obstruent). Based on the onset time of each feature, a multivariate time-series binary matrix (19 features by samples) was produced.
- Phonetic feature onsets (*fo*): univariate vector of the onsets of all phonetic features (i.e., a summary measure of all onsets in *f* without categorical discrimination). This was included as an additional control. Specifically, if the *f* representation above fails to improve EEG prediction beyond the *fo* representation, it would suggest that the encoding of individual phonetic features is not reflected in EEG data.

Model fitting: multivariate linear regression

The acoustic and phonetic representations for both attended and unattended speech described above were normalized and mapped to the concurrently recorded 128-channel EEG signals. Multivariate regularized linear regression was employed to relate the features to the recorded EEG data, where each EEG channel is estimated to be a linear transformation of the speech features over a range of time lags. This transformation is described by the temporal response function (TRF). For a particular speech feature, this operation can be represented mathematically as following (Crosse et al., 2016):

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau)+\epsilon(t,n),$$

where $r(t,n)$ is the neural (EEG) response at channel, n and time point, $t = 1\ldots T$, $s(t-\tau)$ is the multivariate stimulus representation at a lag, τ, $w(\tau,n)$ is the transformation (TRF) of the stimulus at lag τ, and $\epsilon(t,n)$ is the residual response not explained by the model.

The TRF is estimated by minimizing the mean square error between the actual neural response, $r(t,n)$, and the response predicted by the transformation, $\hat{r}(t,n)$. In practice, this can be solved by using reverse correlation. We use the mTRF toolbox (Crosse et al., 2016; https://sourceforge.net/projects/aespa/), which solves for the TRF (w) using reverse correlation with ridge regression:

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

where λ is the ridge regression parameter, I is the identity matrix, and the matrix S is the lagged time series of the stimulus matrix, s. The TRF approach can be used to relate multiple features of the stimuli to the ongoing EEG simultaneously by extending the lagged stimulus matrix to include the various features (more details can be found in Crosse et al., 2016). The ridge regression parameter was tuned using leave-one-out cross validation. That is, for each subject, we trained a separate TRF on each of $n-1$ trials ($n = 40$) for a wide range of λ values (1×10^{-1}, 1 - $\times 10^{0}$, ..., 1×10^{6} for the model comparison analysis, and $2^{-2}, 2^{-1}, \ldots, 2^{37}$ for the partial correlation analysis), computed the average TRF across trials for each λ, then tested the TRFs on the *n*th trial. This was repeated *n* times, rotating the trial to be tested each time. For each subject, the single λ value that maximized the Pearson's correlation coefficient between the actual and predicted neural response averaged over all trials and over all channels was selected. It is worth noting here that the optimal λ parameter value did vary somewhat across subjects. This was not surprising, considering the fact that strength of EEG responses to speech (and hence their signal-to-noise ratio) will vary across subjects because of differences in their cortical fold configuration and skull thickness, etc. However, the optimal λ parameter typically fell into the intermediate region of our tested range, far from the lowest value. This indicates that incorporating regularization improved the model prediction accuracies.

The TRF mapping from stimulus to EEG was computed over a range of time lags, reflecting the idea that changes in the features of the ongoing stimulus are likely to produce effects in the ongoing EEG in that interval. Specifically, we chose the interval 0–300 ms based on previous EEG-based speech studies, where no visible response was present outside this range when considering EEG responses to acoustic and phonetic features (Lalor and Foxe, 2010; Di Liberto et al., 2015). We quantified how well each speech representation related to the neural data using leave-one-trial-out cross-validation (as described above), with Pearson's correlation coefficient as our metric of prediction accuracy. Because the cross-validation procedure takes the average of the validation metric across trials, the models are not biased toward the test data used for cross-validation (Crosse et al., 2016).

To evaluate whether a feature contributed independently of all other features in predicting the neural data, we also computed the partial correlation coefficients (Pearson's *r*) between the EEG predicted by each measure's model with the actual recorded EEG after controlling for the effects of all other features. Specifically, we fit separate cross-validated forward models/TRFs on each of the four speech representations (*s*, *sD*, *f* and *fo*) and predicted EEG based on those models. Then, for the three features to be partialled out (e.g., *s*, *sD*, and *fo*), we used cross-validation to optimally predict our EEG data. We then concatenated these three EEG predictions into one matrix, Z. And, finally, we used the MATLAB built-in function partialcorr (X, Y, Z), where X = the actual recorded EEG, Y = the predicted EEG in response to the feature of interest (the feature whose unique contribution is to be identified, e.g., *f*), and Z = the concatenated predicted EEGs in response to the other features (features that are to be partialled out). This function computes the partial correlation coefficients between X and Y, while controlling for the variables in Z (Fisher, 1924).

Incidentally, we sought to confirm the validity of our partial correlation approach using a second approach. In this approach, we fit one model to three predictors simultaneously using cross-validation (e.g., the *s*, *sD*, and *fo* features). Then we used that model to predict the EEG data and subtracted that predicted data from the real data to leave us with residual EEG. Finally (again using cross validation), we predicted this residual EEG using the left out predictor (e.g., *f*), to see whether this predictor had any power to predict unique variance in the original EEG. This analysis produced the same pattern of result as the approach described in the previous paragraph.

Statistical testing

To test that a partial correlation coefficient is above chance level, we performed nonparametric permutation testing. The predicted EEG activity for each model's representation was permuted across trials such that they were matched to the actual EEG of a different trial, and partial correlation coefficients were computed, controlling for the effects of all other features. This was done 1000 times for each subject to establish a distribution of chance-level prediction accuracies. To perform group-level statistical testing, we generated a null distribution of group means: one prediction accuracy from each subject's individual distribution was selected at random to go into each group mean. This process was repeated 1000 times, sampling with replacement for each subject. For comparison between different models (Fig. 2*C*; Table 1) and for comparison between conditions (attended vs unattended; Fig. 3*A*), two-tailed Wilcoxon signed-rank testing was used. The resulting *p* values were corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). We also used a Bayesian approach to determine the amount of evidence in favor or against the null hypothesis (H_0: no difference). We estimated the Bayes factors (BF_{10}) using the MATLAB toolbox *bayesFactor* toolbox (Moray et al., 2015). More specifically, we used the bf.*t* test function from the toolbox for paired observations that uses Jeffrey–Zellner–Siow (JZS) prior with a scaling factor of 0.707 (Rouder et al., 2012). Typically, any BF_{10} exceeding three is considered to be evidence for the alternative hypothesis (H_1), while below 0.33 is considered strong support for the null hypothesis (H_0), and BF_{10} ranging between 1 and 3 is considered to be weak, anecdotal evidence for the alternative hypothesis (Wagenmakers et al., 2011).

Results

Subjects were found to be compliant in carrying out the behavioral task (Fig. 2*A*). The average questionnaire accuracy was 73.2 ± 3.2% when subjects were tested on the attended story and 27.2 ± 1.6% for unattended stimuli (theoretical chance level is 25% as there are four possible answers to each question).

As mentioned above, our study had two specific questions: (1) does the brain process speech at the level of phonemes and, if so, does it do so for both attended and unattended speech? and (2) at what hierarchical levels does attention influence processing? To address both questions, we chose a two-stage data analysis strategy. The first stage sought to answer the first question. It involved comparing the performance of different models in predicting EEG responses to speech. More specifically, it involved assessing whether or not the inclusion of a categorical phonetic feature representation of speech (*f*) could improve the prediction of EEG beyond that obtained using several other acoustic (*s*, *sD*) and phonetic (*fo*) features. Furthermore, we aimed to assess whether any such improved prediction would hold for both attended and unattended speech. The second analysis was aimed at answering the second question. This strategy involved identifying the ability of each stimulus feature to predict unique variance in the EEG responses (while controlling, or partialling out, the predictive contributions of the other features), and then assessing whether that unique predictive contribution was affected by attention. While these two data analyses are closely related, each is targeted at one of our specific complementary questions.

Model comparisons reveal that responses to attended talkers reflect phonetic-level processing

We assessed the performance of attended and unattended models that were trained to predict EEG responses using the acoustic and phonetic feature representations extracted from the speech stimulus (Fig. 2*B*). The Pearson's correlation between predicted and actual EEG activity, averaged across the same 12 bilateral fronto-temporal channels used in Di Liberto et al. (2015) and Daube et al. (2019), was used as our metric of prediction accuracy for each model (although the same pattern of results was observed if all 128 channels were used). All individual features could predict EEG above chance on a group level (nonparametric permutation test, $p = 0.999e-3$).

Now, the feature spaces used here are not independent, so the individual model predictions are not a pure representation of the extent to which EEG tracks these features. The phonetic features representation (*f*) overlaps with the spectrogram representation (*s*) in that if every phoneme is always spoken the same way, then the two representations would be equivalent. The phoneme onset (*fo*) representation marks the same time-points as the phonetic feature representation (*f*), except that it does not contain feature category information. The *sD*, an approximation of acoustic onsets, contains peaks that overlap with those of the *fo* representation. Thus, the processing of a particular feature was assessed by examining whether adding that feature improved prediction of the EEG responses beyond a model built from other features (Fig. 2*C*). We were particularly interested in establishing whether the three other features could improve prediction accuracy

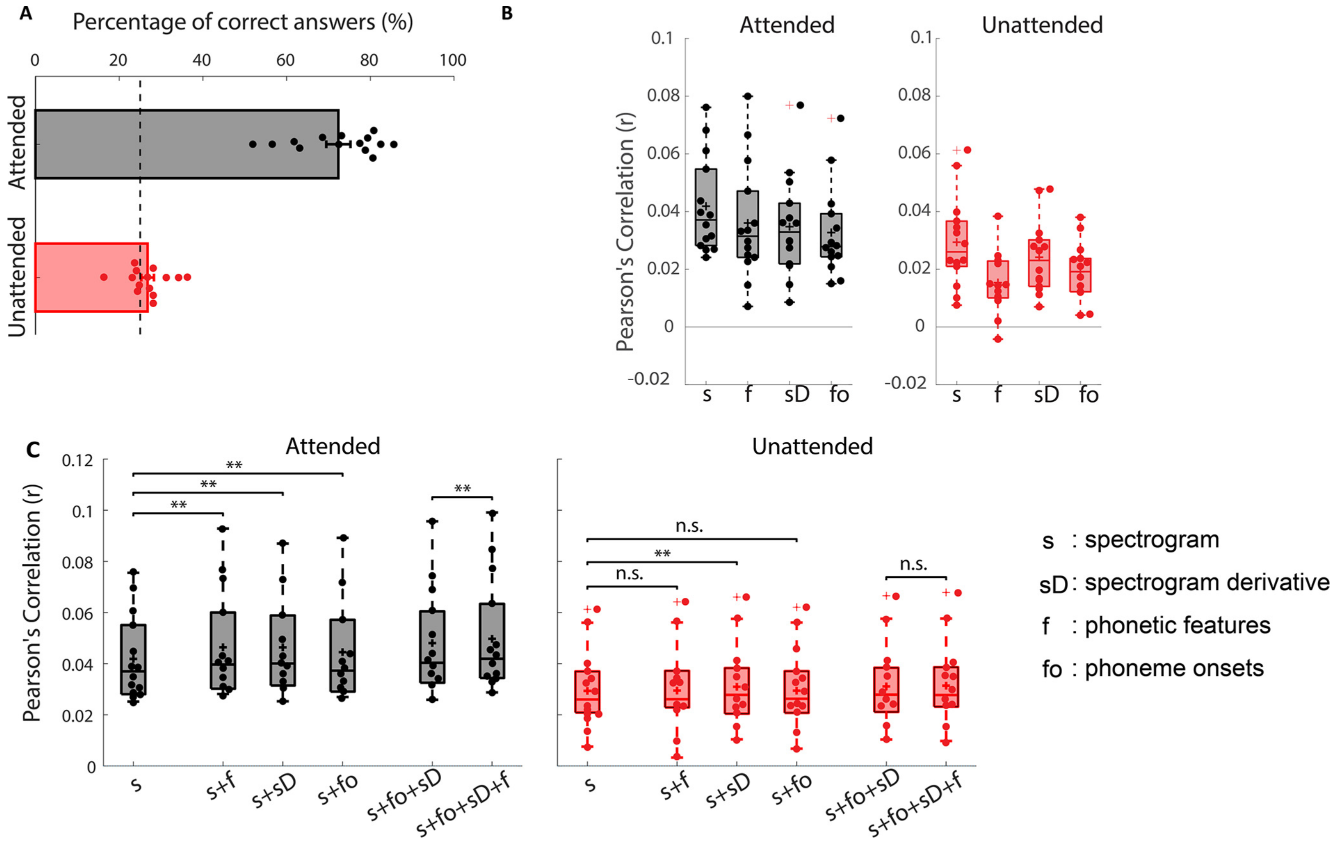

Figure 2. ***A***, Behavioral (comprehension questionnaire) results; dots indicate individual subject performance. Theoretical chance level is 25% (multiple-choice test with four options) and is indicated by the dashed line. ***B***, Prediction accuracies of individual acoustic and phonetic feature spaces (as labeled on horizontal axes) for the attended and unattended stimuli. There is redundancy between these features, so joint modeling was performed to gauge in particular whether phonetic features add unique predictive power beyond other features. ***C***, Prediction accuracies of joint feature spaces for attended and unattended stimuli (**$p < 0.01$, two-tailed Wilcoxon sign-ranked test, FDR corrected). On each box, the central horizontal line indicates the median. The bottom and top edges of the boxes indicate the 25th and 75th percentiles of the data, respectively. The whiskers indicate variability outside the upper and lower quartiles. The '+' sign inside the boxes indicates mean value and the '+' sign outside the boxes indicate outlier. n.s. means not significant ($p > 0.05$).

Table 1. Joint model comparison statistics

	Attended	Unattended
Baseline: *s*		
s + *f* > baseline	**$p = 0.0012$, $z = 3.2330$, $BF_{10} = 8.0159$**	$p = 0.4326$, $z = 0.7847$, $BF_{10} = 0.2790$
s + *sD* > baseline	**$p = 0.0023$, $z = 3.0447$, $BF_{10} = 63.9041$**	**$p = 0.0023$, $z = 3.0047$, $BF_{10} = 13.2765$**
s + *fo* > baseline	**$p = 0.0015$, $z = 3.1708$, $BF_{10} = 2.1335$**	$p = 0.7299$, $z = 0.3453$, $BF_{10} = 0.3035$
Baseline: *s* + *sD*		
s + *sD*+ *f* > baseline	**$p = 0.0043$, $z = 2.8563$, $BF_{10} = 7.1414$**	$p = 0.1981$, $z = 1.2869$, $BF_{10} = 0.6082$
Baseline: *s* + *fo*		
s + *fo* + *f* > baseline	**$p = 0.0023$, $z = 3.0447$, $BF_{10} = 59.7059$**	$p = 0.6378$, $z = 0.4708$, $BF_{10} = 0.2829$
Baseline: *s* + *fo* + *sD*		
s + *fo* + *sD* + *f*> baseline	**$p = 0.0076$, $z = 2.6680$, $BF_{10} = 13.9134$**	$p = 0.4703$, $z = 0.7219$, $BF_{10} = 0.4912$

Wilcoxon two-sided signed-rank test results for attended and unattended stimuli (columns) along with the bayes factors (BF_{10}). Each set of rows tests a different statistical question, as specified by the baseline. Terms being evaluated are to the left of the > symbol. Bolded values indicate significant improvement over baseline.

beyond using spectrograms alone, and if phonetic features contributed unique predictive power beyond all other features. Importantly for comparing the present study with previous studies using this approach (Di Liberto et al., 2015; Daube et al., 2019), our experiment involved speakers with markedly different acoustics (i.e., a male speaker and a female speaker). This means an increase in the variability of the acoustic representation (*s*) for the same phonetic features. As such, the phonetic feature representation should only improve predictions if there are consistent responses to the same phonetic features despite their acoustic variation. In order to test this idea, we performed pairwise statistical testing of the joint model prediction accuracies, these results are shown in Table 1.

In general, we found that adding model features tended to improve EEG prediction performance for attended speech, with fewer improvements visible for unattended speech. Of particular relevance to our original question, we found that including phonetic features (*f*) significantly improved prediction over the acoustic (*s*, *sD*) features when subjects were attending to the speech ($p = 0.0043$), but this improvement was not observed for the unattended speech ($p = 0.1981$). Importantly, the addition of phonetic features also improved prediction when compared with a combination of acoustic features and phoneme onsets (*fo*) in the attended case ($p = 0.0076$), indicating that the inclusion of specific phonetic feature categories carries unique and important information. The acoustic onset (i.e., *sD*) representation displayed a different pattern of results, including this measure along with the spectrogram (i.e., *s*+*sD*) improved prediction accuracy over spectrogram alone for both attended and unattended stimuli, suggesting that it is less strongly modulated by attention. For unattended speech, it is also worth noting that, while the phonetic feature representation did not improve the prediction of EEG responses over

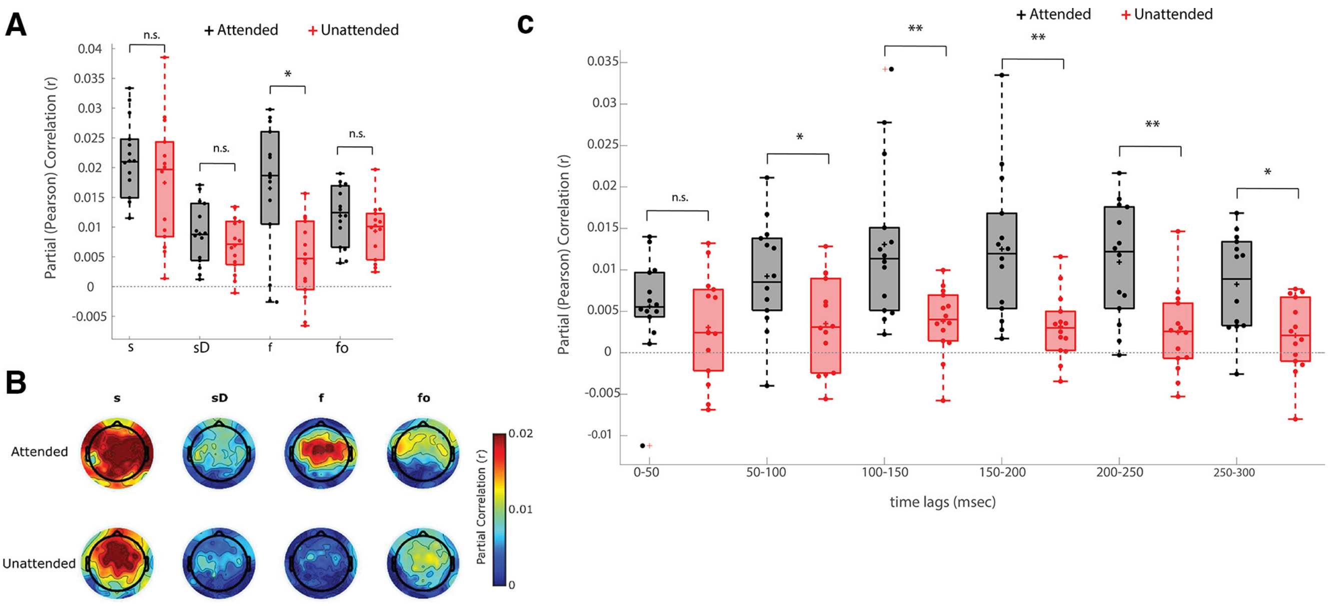

Figure 3. ***A***, Unique predictive power for the acoustic and phonetic feature spaces under two attentional states. The features are as labeled on the horizontal axis. Statistical testing was conducted to identify attentional effects (two-tailed Wilcoxon signed-rank; $*p < 0.05$, FDR corrected). ***B***, Topographic distribution of partial correlations, averaged across all subjects. ***C***, Unique predictive power for the phonetic features representation (*f*) in 50-ms windows from 0 to 300 ms, under two attentional states (two-tailed Wilcoxon signed-rank test; $*p < 0.05$, $**p < 0.005$, FDR corrected). On each box, the central horizontal line indicates the median. The bottom and top edges of the boxes indicate the 25th and 75th percentiles of the data, respectively. The whiskers indicate variability outside the upper and lower quartiles. The '+' sign inside the boxes indicates mean value and the '+' sign outside the boxes indicate outlier. n.s. means not significant ($p > 0.05$).

that based on the acoustic ($s+sD$) or combined acoustic and *fo* ($s+sD+fo$) features, the absence of evidence is not evidence of absence. Indeed, the Bayes factor scores for these two comparisons were $BF_{10} = 0.6082$ and $BF_{10} = 0.4912$, respectively. Both of these are larger than 0.33, indicating we do not have strong evidence in support of the null hypothesis. As such, while we saw no evidence for phonetic feature processing of unattended speech, we cannot claim that it definitely is not happening. However, the evidence in support of such processing for attended speech is strong.

While our results provide evidence for phoneme level processing of attended speech, but no such evidence for unattended speech, the above analysis does not allow us to answer the second question our study set out to examine. In particular, the above results do not allow us to test the hypothesis that attention effects are stronger on measures of phonetic feature processing than on measures of acoustic processing. One might be tempted to conclude from the results in Figure 2*B* that attention is affecting processing at the level of acoustics [compare the prediction accuracies of the *s* model for attended (Fig. 2*B*, left panel) vs unattended (Fig. 2*B*, right panel) speech]. However, as discussed above, the *s* and *f* representations are likely to be highly correlated (given that the acoustics for a given phoneme will tend to be correlated across utterances of that phoneme). As such, if one were to observe an attention effect on the *s* model alone, one would not be able to conclude that attention is affecting low-level acoustic processing. Rather, it could be operating at the level of phonemes and simply appearing in the *s* model performance via correlation. Therefore, to answer our study's second question, we need another approach.

Variance partitioning analysis reveals that attention enhances isolated measures of phonetic feature processing but not acoustic processing

Given the aforementioned redundancies between the predictions of feature spaces, we were also interested in more clearly isolating the unique contribution of each feature. To do so, we employed a partial correlation approach to control for the predictions of all other representations. The unique predictive power of each model is shown in Figure 3*A* (shown here for an average across 12 channels, although including all 128 channels revealed the same pattern of results). On a group level, all features made unique, significant contributions to the EEG predictions except the *sD* models (attended and unattended), and the unattended phonetic features model (nonparametric permutation test; *sD* attended: $p = 0.0619$; *sD* unattended: $p = 0.1788$; f unattended: $p = 0.1788$; all other models: $p < 0.05$). When comparing attended and unattended models, we found a significant effect of attention for phonetic features (two-tailed Wilcoxon signed-rank test; $p < 0.05$, FDR corrected, $BF_{10} = 12.56$), but not for any of the other representations (two-tailed Wilcoxon signed-rank test; FDR corrected; *s*: $p = 0.2412$, $BF_{10} = 0.54$; *sD*: $p = 0.3910$, $BF_{10} = 0.41$; *fo*: $p = 0.1726$, $BF_{10} = 0.59$). Again, the Bayes factor scores for these three latter results are all <1, but >0.33. As such, we cannot definitively claim that there is no effect of attention on the *s*, *sD*, or *fo* representations. However, our data do not provide any evidence of such an effect, unlike the strong effect we see for phonetic features. This strong effect for phonetic features, but not the other representations, is clearly visible in when visualizing the topographic distributions for the unique predictive power of each model (Fig. 3*B*).

We think the above approach of explicitly testing the unique predictive power of different speech representations is a good way to explore the hierarchical processing of natural speech. However, further insight could potentially be gleaned by investigating how the EEG is affected by attention at different temporal delays relative to the speech stimulus. Indeed, in previous work we did this by relating EEG to the speech envelope and finding attention effects at a specific range of time lags from 195 to 230 ms, suggesting that cocktail party attention operates at relatively long latencies, and, thus, beyond the earliest stages of acoustic processing (Power et al., 2012). Similarly, we sought here to explore how the attention effects we have observed on our phonetic features representation might vary across time lags.

To do so, we first explored the possibility of examining the weights of a TRF that captures the unique variance in the EEG because of the phonetic features. To that end, we used cross-validation to fit a three-feature model based on *s*, *sD*, and *fo* to best predict the EEG data. Then we subtracted this EEG prediction from the real EEG, and again used cross validation to model the residual EEG using only the phonetic feature representation. This model was able to predict significant variance in the residual EEG, and these predictions were affected by attention, supporting our earlier results. However, there was no clear effects of attention in the TRF weights themselves at any particular time lags. In fact, this was not terribly surprising. In our previous studies on cocktail party attention, we have found that attention effects on TRF weights themselves are generally much less robust and, thus, require a large amount of data (Power et al., 2012), relative to approaches based on EEG prediction or stimulus reconstruction using those TRFs (O'Sullivan et al., 2015).

With this in mind, we tried a second approach. Specifically, we repeated the approach of predicting EEG using phonetic features *f*, while controlling for *s*, *sD*, and *fo*. However, this time, rather than using a 0- to 300-ms interval of stimulus-response time lags, we ran the analysis separately for different, nonoverlapping 50-ms windows from 0 to 300 ms. Figure 3*C* shows that attention has a significant effect on the phonetic feature-based predictions across a broad range of time lags, starting from around 50 ms. However, the strongest effects of attention are visible from 150 to 250 ms, with weaker effects from 50 to 100, 100 to 150, and from 200 to 250 ms, and no attention effects at the shortest latencies of 0–50 ms (two-tailed Wilcoxon signed-rank test; FDR corrected; 0–50 ms: $p = 0.1726$, $BF_{10} = 0.5468$; 50–100 ms: $p = 0.0166$, $BF_{10} = 3.8610$; 100–150 ms: $p = 0.0017$, $BF_{10} = 11.0430$; 150–200 ms: $p = 6.1035e\text{-}4$, $BF_{10} = 66.8325$; 200–250 ms: $p = 0.0031$, $BF_{10} = 32.4147$; 250–300 ms: $p = 0.0203$, $BF_{10} = 3.6121$). This finding of strongest attention effects at time lags in the range 150–250 ms is largely consistent with earlier results that identified strong cocktail party attention effects at temporal latencies beyond 100 ms with limited effects at earlier latencies (Power et al., 2012; Puvvada and Simon, 2017). One complicating factor here, however, is that exploring EEG predictions using narrow ranges of time delays is not necessarily as highly temporally resolved as it appears to be. This is because of the autocorrelational structure of the speech stimuli and of the EEG. The stimulus-data relationship at, say, 90 ms, can be quite similar to that at 110 ms. So the results need to be interpreted with some sensitivity to the likelihood of smearing in time.

Phonetic feature categories contribute predictive power beyond differentiating vowels and consonants

Di Liberto et al. (2015) found a high degree of discriminability between EEG responses to vowels and consonants. The phonetic feature representation used in the above analysis information on the manner, place, and voicing of consonants, as well as the articulatory position of vowels. We wondered how our results would be affected if, instead of using a 19-dimensional vector, the phonemes were simply marked as vowels or consonants. We repeated the partial correlation analysis described above and found that there was a significant decrease in prediction accuracy when the information on specific articulatory features was left out in the case of attended speech, but not for unattended speech (two-tailed Wilcoxon signed-rank test, $p = 0.011$, $z = 2.542$). Additionally, there was no longer a significant difference between the prediction accuracies of attended and unattended stimuli (att_vc vs unatt_vc; Wilcoxon signed-rank test; $p = 0.4326$; $z = 0.7847$).

Discussion

There has been longstanding debate on how selective attention affects the processing of speech. Here, we set out to investigate attentional modulation at the prelexical level. In particular, based on earlier work (Di Liberto et al., 2015; Daube et al., 2019), we considered how two acoustic and two phonetic feature representations of attended and unattended speech were reflected in EEG responses to that speech. We found that, for attended speech, including a 19-dimensional phonetic features representation improved the prediction of the EEG responses beyond that obtained when only using acoustic features. This was not true for unattended speech. Furthermore, we found that the unique predictive power of the phonetic feature representation was enhanced for attended versus unattended speech. This was not true for any other feature. We contend that these findings make two important contributions to the literature. First, they contribute to the debate around prelexical speech processing in cortex. And second, they contribute to the longstanding debate on how selective attention affects the processing of speech.

In terms of prelexical speech processing, our study shows that a phonetic feature representation has unique predictive power when it comes to modeling responses to attended speech (Fig. 2; Table 1). This suggests that processing attended, but not unattended speech, may involve a mapping from an acoustic to a categorical phonemic representation. This is a controversial idea. In particular, while linguistic theories and prominent speech processing models posit a mapping to phonetic features and/or phonemes as an intermediate stage between low-level acoustics and words (Liberman et al., 1967; McClelland and Elman, 1986; Hickok and Poeppel, 2007), this viewpoint is not unanimous. Some researchers question the necessity and existence of a mental representation at this level for comprehending speech meaning (Lotto and Holt, 2000). Indeed, there have been theories advocating for alternatives, including a different intermediate representational unit (e.g., syllables; Massaro, 1974), or direct matching to the lexicon based on the comparison of acoustic representations (Goldinger, 1998; Klatt, 1989). In support of an acoustic-phonemic stage of processing, there has been evidence from behavioral and EEG studies that humans perceive phoneme categories. Liberman et al. (1957) found that discrimination within phoneme categorical boundaries is poorer than across them. Eimas et al. (1971) looked at infants' ability to discriminate syllables that differed in a voicing (phonetic) feature and found that infants also perceived categorical phonetic features. Additionally, Näätänen (2001) found that the mismatch negativity (MMN), an event-related potential component that reflects discriminable change in some repetitive aspect of the ongoing auditory stimulation, is observed when the deviant stimulus is a phoneme prototype of a subject's native language relative to when it is a nonprototype. However, it has been argued that these studies used simple isolated units, and therefore their results could be because of the type of task and not reflective of everyday listening.

More recent neuroimaging studies have shown that a phonetic feature representation of speech can predict neural activity during listening to natural, continuous speech. Using high-density cortical surface electrodes, Mesgarani and colleagues found that the STG selectively responds to phonetic features (Mesgarani et al., 2014; their Fig. 2–4). Di Liberto et al. (2015) and Khalighinejad et al. (2017) found evidence that noninvasive EEG activity reflects the

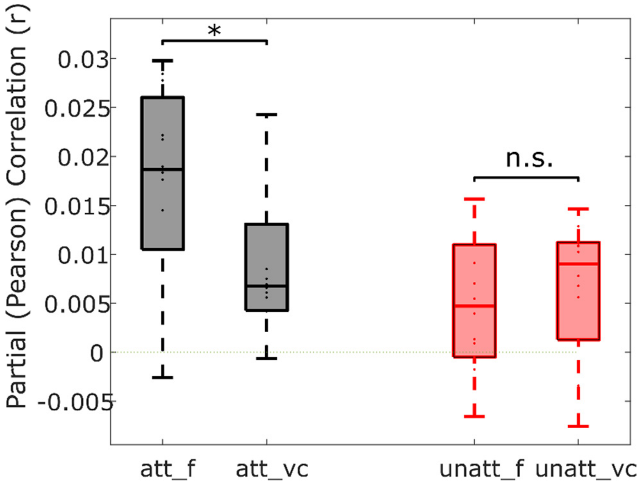

Figure 4. A significant reduction in unique predictive power was observed for the attended condition when the partial correlation analysis was repeated with phonemes only categorized as vowels or consonants (vc), as opposed to their underlying articulatory features (f; two-tailed Wilcoxon signed-rank test, *$p = 0.011$, $z = 2.542$).

categorization of speech into phonetic features. However, a recent paper challenged these findings and argued that neural responses to speech could be well explained by a model that is based entirely on acoustic features (Daube et al., 2019). Namely, they found that acoustic onsets made very similar predictions to the benchmark phonetic features, and that acoustic onsets explained parts of the neural response that phonetic features could not. Here, we show evidence to support the encoding of a rich array of phonetic features (certainly more than simply vowel and consonant categorization; Fig. 4) for attended speech. It is possible that we were able to find unique predictive power for these phonetic features because our study involved stimuli from two different talkers (a male and a female) while the stimuli used in Daube et al. (2019) were spoken by a single talker. As mentioned earlier, there is redundancy between speech acoustics and a phonetic feature representation because each phonetic feature has a characteristic spectro-temporal acoustic profile. This overlap can make it difficult to isolate the precise contribution of each predictor. Indeed, if there were no variation in spectro-temporal acoustics across each utterance of the same phoneme, the phonetic feature and spectrogram representations would be, effectively, identical. However, the more variation there is between utterances of the same phoneme (which will happen when including both male and female speakers), the greater the difference should be between the contribution of neurons that care about those variations (e.g., low-level "acoustic" neurons) and the contribution of neurons whose responses are invariant to the acoustic details and only dependent on phoneme category. As such, it may be that, when using stimuli with high intra-phoneme acoustic variation, including both acoustic and phonetic feature representations is of most benefit for predicting neural activity. Future work will investigate this idea further.

Interestingly, while our data support the idea of a mapping from an acoustic to a categorical phonemic representation during attended speech processing, we found no such evidence for unattended speech. Our Bayesian analysis ($BF_{10} = 0.4912 > 0.33$, last row of Table 1) does not allow us to confidently assert that no such mapping takes place. For example, it could be that the generally lower prediction scores for unattended speech means that the signal-to-noise ratio of the data are just not high enough to detect subtle effects. That said, the EEG prediction scores for our individual (Fig. 2*B*) and combined (Fig. 2*C*) speech representations for unattended speech were all significantly above chance. Furthermore, the unique predictive power of the phonetic feature model for unattended speech was very low (Fig. 3*A*,*B*). As such, it seems clear that, if it takes place at all, categorical processing of phonetic features for unattended speech is much reduced. Importantly, we contend that the lack of any evidence for phonetic feature processing of unattended speech actually further supports our claim that such phonetic feature processing is occurring for attended speech. In particular, if it were true that putative categorical phonemic responses to speech could be well explained by a model based entirely on acoustic features, then this would surely hold for both attended and unattended speech. That is, unless some acoustic features are more strongly modulated by attention than others. But we favor the interpretation that categorical linguistic processing (which surely happens) occurs for attended and not unattended speech and that this dichotomy extends from semantic (Broderick et al., 2018) and other linguistic levels (Brodbeck et al., 2018) down to the level of phoneme categorization.

As a final comment on this issue, we acknowledge that there are also other acoustic transformations of the auditory stimulus that have not been considered beyond the spectrogram and its derivative, as well as other theorized intermediate representations apart from phonemes/phonetic features that have not been tested. It is likely that some of these feature sets may overlap with our measure of phonetic features. As such, we cannot definitively prove the encoding of this measure in cortex. However, one reason the idea of mapping to words via an acoustic-phonetic stage has prevailed is because of its computational efficiency (Scharenborg et al., 2005). It is less obvious how the detection of acoustic onsets and other acoustic representations would lead to lexical access without an intermediate stage – and in particular, how such a model would robustly generalize to new words and new speakers.

As to the cocktail party attention debate, our results suggest that attention differentially modulates cortical processing of acoustic and phonetic information. We found that the *unique* predictive power of phonetic features was modulated by attention, while the other feature spaces were not. This result is in line with a wealth of evidence for stronger attention effects at higher levels of the speech processing hierarchy. For example, there is no doubt that both attended and unattended speech will influence activity in the cochlea and the auditory nerve. But by the time one reaches auditory association areas like STG, only the attended speaker is represented (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013). Indeed, an important recent study (O'Sullivan et al., 2019) used simultaneous recordings in both primary auditory cortex (Heschl's gyrus; HG) and STG, to show that both attended and unattended speech are robustly represented in primary areas, with STG selectively representing the attended speaker. This lines up very well with our results. In particular, it is well established that spectro-temporal models of sounds can do a good job of explaining neural responses in primary auditory areas, but not in nonprimary auditory cortex (Norman-Haignere and McDermott, 2018). Thus, we might expect our isolated acoustic speech processing indices to mostly reflect activity from primary auditory areas, which have been shown to be relatively unaffected by attention (O'Sullivan et al., 2019). Meanwhile STG is known to be selective for phonetic feature representations (Mesgarani et al., 2014) and activity in STG is strongly affected by attention (O'Sullivan et al., 2019). Thus, we might expect our isolated measures of phonetic feature processing to be strongly influenced by activity from STG, and thus, significantly affected by attention. This is the pattern of results that we have observed.

Additional support for our findings as evidence for a higher-level locus of attentional selection comes from research examining the latency of attention effects on EEG and MEG responses to speech. This approach has commonly focused on a rather general measure of speech processing that derives from indexing how the neural data track the speech envelope. Directly determining how much this speech processing is driven by acoustic versus speech-specific processing is not clear, although there is undoubtedly a very substantial acoustic contribution (Lalor et al., 2009; Howard and Poeppel, 2010; Prinsloo and Lalor, 2020). Nonetheless, researchers have examined components of these responses at different latencies as a proxy measure of processing at different hierarchical stages, and found that both attended and unattended speech are well represented in early components, with distinct responses to the attended speaker appearing only in late components (Ding and Simon, 2012; Power et al., 2012; Puvvada and Simon, 2017). Directly linking envelope tracking at different latencies with different representations of speech is a task for future research. Indeed, it may be quite an important task given the myriad efforts in recent years aimed at linking the effects of attention on behavior with those on a difficult-to-interpret neurophysiological measure that is likely substantially driven by attentionally-insensitive acoustic processing (Tune et al., 2020). Similarly, it may be that decoding of attentional selection from neural recordings will be improved by focusing on how those recordings reflect specific speech representations that are strongly affected by attention rather than the very general speech envelope (O'Sullivan et al., 2015).

Finally, it is important to add the caveat that our lack of an attention effect on isolated measures of acoustic processing is not evidence of their absence. Indeed, while the Bayes factor scores for the effect of attention on the unique predictive power of the *s* and *sD* models were both <1, they were not <0.33, meaning we cannot definitively assert that there are no attention effects at the acoustic level. It may well be that our analysis pipeline is not sensitive enough to detect small attention effects on those measures. Nonetheless, the fact that a clear difference was found for the phonetic feature representation supports the notion of stronger attention effects at higher levels of the speech processing hierarchy. The capacity of the brain to process information from multiple streams at any given moment in time is limited, so a longstanding question has concerned the extent to which unattended speech undergoes processing in cortex. Our results, considered together with studies focused on lexical (Brodbeck et al., 2018) and semantic (Broderick et al., 2018) processing, suggest it possible that attention is categorically selective for speaker-invariant representations and, at most, attenuates lower-level acoustic (speaker-dependent) measures.

References

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

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The CIPIC HRTF database. Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575), pp 99–102.

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.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512–528.

Broadbent D (1958) Perception and communication. London: Pergamon.

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.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979.

Chomsky N, Halle M (1968) The sound pattern of English. New York: Harper and Row.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The multivariate temporal response function (MTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604.

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.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423–3431.

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.

Deutsch JA, Deutsch D (1963) Some theoretical considerations. Psychol Rev 70:80–90.

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.

Di Liberto GM, ong D, Melnik GA, de Cheveigné A (2019) Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196:237–247.

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89.

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J (1971) Speech perception in infants. Science 171:303–306.

Fisher RA (1924) The distribution of the partial correlation coefficient. Metron 3:329–332.

Goldinger SD (1998) Echoes of echoes? An episodic theory of lexical access. Psychol Rev 105:251–279.

Gwilliams L, King J-R, Marantz A, Poeppel D (2020) Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. bioRxiv 2020.04.04.025684.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Holdgraf CR, Rieger JW, Micheli C, Martin S, Knight RT, Theunissen FE (2017) Encoding and decoding models in cognitive electrophysiology. Front Syst Neurosci 11:61.

Howard MF, Poeppel D (2010) Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J Neurophysiol 104:2500–2511.

Johnston WA, Wilson J (1980) Perceptual processing of nontargets in an attention task. Mem Cognit 8:372–377.

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98:630–644.e16.

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.

Klatt DH (1989) Review of selected models of speech perception. In: Lexical representation and process, pp 169–226. Cambridge: The MIT Press.

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31:189–193.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 102:349–359.

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54:358–368.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461.

Lotto AJ, Holt L (2000) The Illusion of the phoneme. In: (Billing SJ, Boyle JP, and Griffith AM, eds), Chicago Linguistic Society, Volume 35: The Panels. (pp. 191–204). Chicago Linguistic Society: Chicago.

Massaro DW (1974) Perceptual units in speech recognition. J Exp Psychol 102:199–208.

McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M (2017) Montreal forced aligner: trainable text-speech alignment using Kaldi. Proc. Interspeech, pp. 498–502. Available at doi:10.21437/Interspeech. 2017-1386.

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010.

Moray N (1959) Attention in dichotic listening: affective cues and the influence of instructions. Q J Exp Psychol 11:56–60.

Moray RD, Rouder JN, Jamil T (2015) BayesFactor: Computation of Bayes factors for common designs. R package version 0.9. 12-2.

Näätänen R (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38:1–21.

Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol 16:e2005127.

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20:2486–2495.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706.

O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM, Sheth SA, Mehta AD, Mesgarani N (2019) Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron 104:1195–1209.e3.

Peelle JE, Johnsrude I, Davis MH (2010) Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4:51.

Power AJ, Lalor EC, Reilly RB (2011) Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. Cereb Cortex 21:1223–1230.

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

Prinsloo KD, Lalor EC (2020) General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras. bioRxiv 348557. doi: 10.1101/2020.10.21.348557.

Puvvada KC, Simon JZ (2017) Cortical representations of speech in a multi-talker auditory scene. J Neurosci 37:9189–9196.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724.

Reddy S, Stanford JN (2015) Toward completely automated vowel extraction: introducing DARLA. Ling Vang 1:15–28.

Rosenfelder IJ, Fruehwald K, Evanini S, Seyfarth K, Gorman H, Prichard J Yuan (2014) FAVE (forced alignment and vowel extraction) suite version 1.1.3. Available at https://doi.org/10.5281/ZENODO.9846.

Rouder JN, Morey RD, Speckman PL, Province JM (2012) Default Bayes factors for ANOVA designs. J Math Psychol 56:356–374.

Schädler MR, Kollmeier B (2015) Separable spectro-temporal Gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. J Acoust Soc Am 137:2047–2059.

Scharenborg O, Norris D, Bosch L, McQueen JM (2005) How should a speech recognizer work? Cogn Sci 29:867–918.

Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. Science 357:797–801.

Teoh ES, Lalor EC (2019) EEG decoding of the target speaker in a cocktail party scenario: considerations regarding dynamic switching of talker location. J Neural Eng 16:036017.

Teoh ES, Cappelloni MS, Lalor EC (2019) Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. Eur J Neurosci 50:3831–3842.

Treisman AM (1964) Verbal cues, language, and meaning in selective attention. Am J Psychol 77:206–219.

Tune S, Alavash M, Fiedler L, Obleser J (2020) Neural attention filters do not predict behavioral success in a large cohort of aging listeners. bioRxiv 2020.05.20.105874.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HLJ (2011) Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). J Pers Soc Psychol 100:426–432.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. Neuron 77:980–991.