JNeurosci Accepted Manuscript

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment

Emily S. Teoh[1], Farhin Ahmed[2], 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

[2] Department of Neuroscience, Department of Biomedical Engineering, and Del Monte Neuroscience Institute, University of Rochester, Rochester 14627, NY, USA

Corresponding author email address: Edmund_Lalor@urmc.rochester.edu

Running title: Attention enhances phonetic feature encoding

Abstract

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 pre-lexical representations of speech, insights that complement recent anatomical 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.

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 not for acoustic

38 processing. These findings add an important new layer in our understanding of how the human brain
39 solves the cocktail party problem.

40 **Introduction**

41 The ability to focus on a single talker amidst multiple sounds is essential for human
42 communication. Cherry (1953) brought this phenomenon to the fore, investigating our behavioral
43 capacity for selective auditory attention. His study stimulated subsequent inquiries and the
44 proposal of various theories of attention (Broadbent 1958; Moray 1959; Deutsch and Deutsch
45 1963; Treisman 1964; Johnston and Wilson 1980). These theories model attention as a selective
46 filter that rejects the unattended message beyond a certain information processing stage.
47 However, whether this filter operates at an early or late stage of speech processing is still
48 unresolved.

49 In recent years, neuroscience has sought to contribute to this debate by leveraging our increased
50 understanding of the hierarchical nature of speech processing in the human cortex. In particular,
51 to construe meaning from sound, the cortex is posited to compute several intermediate levels of
52 increasingly abstract representations in different functionally specialized regions of a
53 hierarchically organized cortical network (Hickok and Poeppel 2007; Peelle et al. 2010;
54 Rauschecker and Scott 2009; DeWitt and Rauschecker 2012; Kell et al. 2018). For example,
55 single talker neuroimaging studies have revealed that core auditory regions code low-level
56 features which are combined in higher areas to yield more abstract neural codes (Binder et al.
57 2000; Davis and Johnsrude 2003; Okada et al. 2010). And in terms of multi-talker selective
58 attention, studies have shown that primary auditory cortical responses represent all talkers
59 irrespective of attentional state (J. O'Sullivan et al. 2019), but in "higher" areas such as the
60 superior temporal gyrus (STG), only the attended speaker is represented (Mesgarani and Chang
61 2012; Zion Golumbic et al. 2013; J. O'Sullivan et al. 2019). Additionally, EEG/MEG studies
62 examining the latency of neural responses (which can be considered a rough proxy measure of
63 processing at different hierarchical stages) have shown that all talkers are co-represented in early
64 components, with distinct responses to the attended speaker appearing only in later components
65 (Ding and Simon 2012b; Power et al. 2012; Puvvada and Simon 2017). Importantly, this
66 specificity is not something that is necessarily true for more simplistic non-speech stimuli and
67 tasks (e.g., Power et al., 2011).

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

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

99

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 neurological disorder. Each subject provided written informed consent prior to 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 one-minute 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) & "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 degrees 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).

128 All stimuli were sampled with a frequency of 44 100 Hz and were presented using Sennheiser

129 HD650 headphones and Presentation software from Neurobehavioral Systems

130 (http://www.neurobs.com).

131 Data Acquisition and Pre-processing

132 The experiment was conducted in a soundproof room. A Biosemi ActiveTwo system was used to

133 record EEG data from 128 electrode positions on the scalp as well as two electrodes over the

134 mastoid processes (all digitized at 512Hz).

135 EEG data were re-referenced to the mastoids. Automatic bad channel rejection and interpolation

136 was performed. A particular channel was deemed as bad if the standard deviation of the channel

137 was lower than a third of or exceeded three times the mean of the standard deviation of all

138 channels. In place of the bad channel, data were interpolated from the four nearest neighboring

139 electrodes using spherical spline interpolation (Delorme and Makeig 2004). To decrease

140 subsequent processing time, data were downsampled to 128Hz. Data were filtered between 0.2 –

141 8 Hz using a zero-phase shift FIR filter with Kaiser window (filter order was 20000 for the 0.2

142 Hz high pass filter and 1000 for the 8 Hz low pass filter; maximum stopband attenuation was -60

143 dB). Low delta-band frequencies (down to 0.2 Hz) were included as they have previously been

144 found to be important for speech processing (Teoh et al, 2019).

145 Speech Representations

146 We computed acoustic and phonetic speech representations of both the attended and unattended

147 stories (Figure 1A depicts these representations for an excerpt taken from one of the audio clips).

148 Acoustic:

149 • Spectrogram (s): The GBFB toolbox (Schädler and Kollmeier 2015) was used to extract

150   the spectral decomposition of the time-varying stimulus energy in 31 mel-spaced bands

151   with logarithmic compressive nonlinearity. The spectrogram representation computed in

152   this way is consistent with (Daube et al., 2019), where it was shown to better predict

153   neural activity than other techniques.

154 • Spectrogram Derivative (sD): The temporal derivative of each spectrogram channel was

155   computed, and then half-wave rectified. This represents an approximation of the onsets in

156 the acoustics and was shown to contribute to predicting neural responses in previous

157 research (Daube et al., 2019).

158 Phonetic:

159 • Phonetic features (*f*): Phoneme-level segmentation of the stimuli was first computed

160 using FAVE-Extract (Rosenfelder et al. 2014) and the Montreal Forced Aligner

161 (McAuliffe et al. 2017) via the DARLA web interface (Reddy and Stanford 2015). The

162 phoneme representation was then mapped into a space of 19 features based on the

163 mapping described previously (Chomsky and Halle 1968; Mesgarani et al. 2014). The

164 features describe the articulatory and acoustic properties of the phonetic content of

165 speech (e.g., bilabial, plosive, voiced, obstruent). Based on the onset time of each feature,

166 a multivariate time-series binary matrix (19 features by samples) was produced.

167 • Phonetic feature onsets (*fo*): Univariate vector of the onsets of all phonetic features (i.e., a

168 summary measure of all onsets in *f* without categorical discrimination). This was

169 included as an additional control. Specifically, if the *f* representation above fails to

170 improve EEG prediction beyond the *fo* representation, it would suggest that the encoding

171 of individual phonetic features is not reflected in EEG data.

172

173

174 <u>Model fitting: Multivariate Linear Regression</u>

175 The acoustic and phonetic representations for both attended and unattended speech described

176 above were normalized and mapped to the concurrently recorded 128-channel EEG signals.

177 Multivariate regularized linear regression was employed to relate the features to the recorded

178 EEG data, where each EEG channel is estimated to be a linear transformation of the speech

179 features over a range of time lags. This transformation is described by the temporal response

180 function (TRF). For a particular speech feature, this operation can be represented mathematically

181 as following (Crosse et al. 2016):

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

6

182   where $r(t,n)$ is the neural (EEG) response at channel, n and time point, $t = 1 \dots T$, $s(t - \tau)$ is

183   the multivariate stimulus representation at a lag, τ, $w(\tau,n)$ is the transformation (TRF) of the

184   stimulus at lag τ, and $\epsilon(t,n)$ is the residual response not explained by the model.

185   The TRF is estimated by minimizing the mean square error between the actual neural response,

186    $r(t,n)$, and the response predicted by the transformation, $\hat{r}(t,n)$. In practice, this can be solved

187   by using reverse correlation. We use the mTRF toolbox (Crosse et al. 2016-

188   https://sourceforge.net/projects/aespa/), which solves for the TRF (w) using reverse correlation

189   with ridge regression:

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

190   where λ is the ridge regression parameter, I is the identity matrix, and the matrix S is the lagged

191   time series of the stimulus matrix, s. The TRF approach can be used to relate multiple features of

192   the stimuli to the ongoing EEG simultaneously by extending the lagged stimulus matrix to

193   include the various features (more details can be found in Crosse et al. 2016). The ridge

194   regression parameter was tuned using leave-one-out cross validation. That is, for each subject,

195   we trained a separate TRF on each of n-1 trials (n = 40) for a wide range of λ values ($1\text{x}10^{-1}$,

196    $1\text{x}10^{0}$, …, $1\text{x}10^{6}$ for the model comparison analysis, and 2^{-2}, 2^{-1}, …, 2^{37} for the partial correlation

197   analysis), computed the average TRF across trials for each λ, then tested the TRFs on the nth

198   trial. This was repeated n times, rotating the trial to be tested each time. For each subject, the

199   single λ value that maximized the Pearson's correlation coefficient between the actual and

200   predicted neural response averaged over all trials and over all channels was selected. It is worth

201   noting here that the optimal λ parameter value did vary somewhat across subjects. This was not

202   surprising, considering the fact that strength of EEG responses to speech (and hence their signal-

203   to-noise ratio) will vary across subjects due to differences in their cortical fold configuration and

204   skull thickness, etc. However, the optimal λ parameter typically fell into the intermediate region

205   of our tested range, far from the lowest value. This indicates that incorporating regularization

206   improved the model prediction accuracies.

207   The TRF mapping from stimulus to EEG was computed over a range of time lags, reflecting the

208   idea that changes in the features of the ongoing stimulus are likely to produce effects in the

209   ongoing EEG in that interval. Specifically, we chose the interval 0 to 300 ms based on previous

210   EEG-based speech studies, where no visible response was present outside this range when

211   considering EEG responses to acoustic and phonetic features (Di Liberto et al., 2015; Lalor and

212   Foxe 2010). We quantified how well each speech representation related to the neural data using

213   leave-one-trial-out cross-validation (as described above), with Pearson's correlation coefficient

214   as our metric of prediction accuracy. Because the cross-validation procedure takes the average of

215   the validation metric across trials, the models are not biased toward the test data used for cross-

216   validation (Crosse et al. 2016).

217   To evaluate whether a feature contributed independently of all other features in predicting the

218   neural data, we also computed the partial correlation coefficients (Pearson's r) between the EEG

219   predicted by each measure's model with the actual recorded EEG after controlling for the effects

220   of all other features. Specifically, we fit separate cross-validated forward models/TRFs on each

221   of the four speech representations (s, sD, f and fo) and predicted EEG based on those models.

222   Then, for the three features to be partialled out (e.g., s, sD, and fo), we used cross-validation to

223   optimally predict our EEG data. We then concatenated these three EEG predictions into one

224   matrix, Z. And, finally, we used the MATLAB built-in function partialcorr (X, Y, Z) where X =

225   the actual recorded EEG, Y = the predicted EEG in response to the feature of interest (the feature

226   whose unique contribution is to be identified, e.g., f), and Z = the concatenated predicted EEGs

227   in response to the other features (features that are to be partialled out). This function computes

228   the partial correlation coefficients between X and Y, while controlling for the variables in Z

229   (Fisher 1924).

230   Incidentally, we sought to confirm the validity of our partial correlation approach using a second

231   approach. In this approach, we fit one model to three predictors *simultaneously* using cross-

232   validation (e.g., the s, sD and fo features). Then we used that model to predict the EEG data and

233   subtracted that predicted data from the real data to leave us with residual EEG. Finally, (again

234   using cross validation), we predicted this residual EEG using the left out predictor (e.g., f), to see

235   if this predictor had any power to predict unique variance in the original EEG. This analysis

236   produced the same pattern of result as the approach described in the previous paragraph.

237   <u>Statistical Testing</u>

238   To test that a partial correlation coefficient is above chance level, we performed non-parametric

239   permutation testing. The predicted EEG activity for each model's representation was permuted

8

240 across trials such that they were matched to the actual EEG of a different trial, and partial

241 correlation coefficients were computed, controlling for the effects of all other features. This was

242 done 1000 times for each subject to establish a distribution of chance-level prediction accuracies.

243 To perform group-level statistical testing, we generated a null distribution of group means: one

244 prediction accuracy from each subject's individual distribution was selected at random to go into

245 each group mean. This process was repeated 1000 times, sampling with replacement for each

246 subject. For comparison between different models (Figure 2C and Table 1) and for comparison

247 between conditions (attended vs unattended, Figure 3A), two-tailed Wilcoxon signed-rank

248 testing was used. The resulting p values were corrected using the False Discovery Rate (FDR)

249 method (Benjamini & Hochberg, 1995). We also used a Bayesian approach to determine the

250 amount of evidence in favor or against the null hypothesis (H_0: no difference). We estimated the

251 Bayes factors (BF_{10}) using the MATLAB toolbox *bayesFactor* toolbox (Morey et al., 2015).

252 More specifically, we used the bf.ttest function from the toolbox for paired observations that uses

253 Jeffrey-Zellner-Siow (JZS) prior with a scaling factor of 0.707 (Rouder et al. 2012). Typically,

254 any BF_{10} exceeding 3 is considered to be evidence for the alternative hypothesis (H_1), while

255 below 0.33 is considered strong support for the null hypothesis (H_0), and BF_{10} ranging between

256 1-3 is considered to be weak, anecdotal evidence for the alternative hypothesis (Wagenmakers et

257 al. 2011).

258 **[Fig 1 about here]**

259 Figure 1: (A) Speech representations: The first row depicts an acoustic waveform of an excerpt

260 taken from one of the stimuli. Subsequent rows show the computed acoustic and phonetic

261 representations for that excerpt. (B) Analysis framework: Cross-validation is used to train

262 forward models mapping the different attended and unattended speech representations to EEG.

263 These models are then used to predict left-out EEG. Pearson's correlation is used to evaluate

264 model accuracy, and the prediction accuracies of acoustic and phonetic models are compared.

265

266 **Results**

267 Subjects were found to be compliant in carrying out the behavioral task (Figure 2A). The average

268 questionnaire accuracy was $73.2 \pm 3.2\%$ when subjects were tested on the attended story and

269 $27.2 \pm 1.6\%$ for unattended stimuli (theoretical chance level is 25% as there are four possible

270 answers to each question).

271 As mentioned above, our study had two specific questions: 1) does the brain process speech at

272 the level of phonemes and, if so, does it do so for both attended and unattended speech? And, 2)

273 at what hierarchical levels does attention influence processing? To address both questions, we

274 chose a two-stage data analysis strategy. The first stage sought to answer the first question. It

275 involved comparing the performance of different models in predicting EEG responses to speech.

276 More specifically, it involved assessing whether or not the inclusion of a categorical phonetic

277 feature representation of speech (*f*) could improve the prediction of EEG beyond that obtained

278 using several other acoustic (*s*, *sD*) and phonetic (*fo*) features. Furthermore, we aimed to assess

279 whether any such improved prediction would hold for both attended and unattended speech. The

280 second analysis was aimed at answering the second question. This strategy involved identifying

281 the ability of each stimulus feature to predict unique variance in the EEG responses (while

282 controlling, or partialling out, the predictive contributions of the other features), and then

283 assessing whether that unique predictive contribution was affected by attention. While these two

284 data analyses are closely related, each is targeted at one of our specific complementary questions.

285 <u>Model comparisons reveal that responses to attended talkers reflect phonetic-level processing</u>

286 We assessed the performance of attended and unattended models that were trained to predict

287 EEG responses using the acoustic and phonetic feature representations extracted from the speech

288 stimulus (Figure 2B). The Pearson's correlation between predicted and actual EEG activity,

289 averaged across the same 12 bilateral fronto-temporal channels used in Di Liberto et al., (2015)

290 and Daube et al. (2019), was used as our metric of prediction accuracy for each model (although

291 the same pattern of results was observed if all 128 channels were used). All individual features

292 could predict EEG above chance on a group level (non-parametric permutation test, p=0.999e-3).

293 Now, the feature spaces used here are not independent, so the individual model predictions are

294 not a pure representation of the extent to which EEG tracks these features. The phonetic features

10

295 representation (*f*) overlaps with the spectrogram representation (*s*) in that if every phoneme is

296 always spoken the same way, then the two representations would be equivalent. The phoneme

297 onset (*fo*) representation marks the same time-points as the phonetic feature representation (*f*),

298 except that it does not contain feature category information. The spectrogram derivative (*sD*), an

299 approximation of acoustic onsets, contains peaks that overlap with those of the phoneme onset

300 (*fo*) representation. Thus, the processing of a particular feature was assessed by examining

301 whether adding that feature improved prediction of the EEG responses beyond a model built

302 from other features (Figure 2C). We were particularly interested in establishing whether the three

303 other features could improve prediction accuracy beyond using spectrograms alone, and if

304 phonetic features contributed unique predictive power beyond all other features. Importantly for

305 comparing the present study with previous studies using this approach (Di Liberto et al., 2015;

306 Daube et al., 2019), our experiment involved speakers with markedly different acoustics (i.e., a

307 male speaker and a female speaker). This means an increase in the variability of the acoustic

308 representation (*s*) for the same phonetic features. As such, the phonetic feature representation

309 should only improve predictions if there are consistent responses to the same phonetic features

310 *despite* their acoustic variation. In order to test this idea, we performed pairwise statistical testing

311 of the joint model prediction accuracies – these results are shown in Table 1.

312 **[Fig 2 about here]**

313

314 Figure 2 | (A) Behavioral (comprehension questionnaire) results – dots indicate individual

315 subject performance. Theoretical chance level is 25% (multiple-choice test with four options)

316 and is indicated by the dashed line (B) Prediction accuracies of individual acoustic and phonetic

317 feature spaces (as labelled on horizontal axes) for the attended and unattended stimuli. There is

318 redundancy between these features, so joint modelling was performed to gauge in particular

319 whether phonetic features adds unique predictive power beyond other features (C) Prediction

320 accuracies of joint feature spaces for attended and unattended stimuli (**$p<0.01$, two-tailed

321 Wilcoxon sign-ranked test, FDR corrected).

322

323

324 Table 1 | Joint model comparison statistics: Wilcoxon two-sided signed rank test results for
325 attended and unattended stimuli (columns) along with the Bayes factors (BF_{10}). Each set of rows
326 tests a different statistical question, as specified by the baseline. Terms being evaluated are to the
327 left of the '>' symbol. Bolded values indicate significant improvement over baseline.

	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

328

329

330 In general, we found that adding model features tended to improve EEG prediction performance
331 for attended speech, with fewer improvements visible for unattended speech. Of particular
332 relevance to our original question, we found that including phonetic features (*f*) significantly
333 improved prediction over the acoustic (*s*, *sD*) features when subjects were attending to the
334 speech (*p* = 0.0043); but this improvement was not observed for the unattended speech (*p* =
335 0.1981). Importantly, the addition of phonetic features also improved prediction when compared
336 with a combination of acoustic features and phoneme onsets (*fo*) in the attended case (p =

337 0.0076), indicating that the inclusion of specific phonetic feature categories carries unique and

338 important information. The acoustic onset (i.e., spectrogram derivative) representation displayed

339 a different pattern of results – including this measure along with the spectrogram (i.e., $s+sD$)

340 improved prediction accuracy over spectrogram alone for both attended and unattended stimuli,

341 suggesting that it is less strongly modulated by attention. For unattended speech, it is also worth

342 noting that, while the phonetic feature representation did not improve the prediction of EEG

343 responses over that based on the acoustic ($s+sD$) or combined acoustic and phoneme onset

344 ($s+sD+fo$) features, the absence of evidence is not evidence of absence. Indeed, the Bayes factor

345 scores for these two comparisons were $BF_{10} = 0.6082$ and $BF_{10} = 0.4912$, respectively. Both of

346 these are larger than 0.33, indicating we do not have strong evidence in support of the null

347 hypothesis. As such, while we saw no evidence for phonetic feature processing of unattended

348 speech, we cannot claim that it definitely is not happening. However, the evidence in support of

349 such processing for attended speech is strong.

350 While our results provide evidence for phoneme level processing of attended speech, but no such

351 evidence for unattended speech, the above analysis does not allow us to answer the second

352 question our study set out to examine. In particular, the above results do not allow us to test the

353 hypothesis that attention effects are stronger on measures of phonetic feature processing than on

354 measures of acoustic processing. One might be tempted to conclude from the results in Fig 2B

355 that attention is affecting processing at the level of acoustics (compare the prediction accuracies

356 of the s model for attended [left panel] vs unattended [right panel] speech). However, as

357 discussed above, the s and f representations are likely to be highly correlated (given that the

358 acoustics for a given phoneme will tend to be correlated across utterances of that phoneme). As

359 such, if one were to observe an attention effect on the s model alone, one would not be able to

360 conclude that attention is affecting low-level acoustic processing. Rather, it could be operating at

361 the level of phonemes and simply appearing in the s model performance via correlation.

362 Therefore, in order to answer our study's second question, we need another approach.

363 <u>Variance partitioning analysis reveals that attention enhances isolated measures of phonetic</u>
364 <u>feature processing, but not acoustic processing</u>

365 Given the aforementioned redundancies between the predictions of feature spaces, we were also

366 interested in more clearly isolating the unique contribution of each feature. To do so, we

13

367   employed a partial correlation approach to control for the predictions of all other representations.

368   The unique predictive power of each model is shown in Figure 3A (shown here for an average

369   across 12 channels, although including all 128 channels revealed the same pattern of results). On

370   a group level, all features made unique, significant contributions to the EEG predictions except

371   the spectrogram derivative models (attended and unattended), and the unattended phonetic

372   features model (non-parametric permutation test; *sD* attended: p=0.0619; *sD* unattended:

373   p=0.1788; *f* unattended: p=0.1788; all other models: p<0.05). When comparing attended and

374   unattended models, we found a significant effect of attention for phonetic features (two-tailed

375   Wilcoxon signed rank test; p<0.05, FDR corrected, $BF_{10} = 12.56$), but not for any of the other

376   representations (two-tailed Wilcoxon signed rank test; FDR corrected; *s*: p = 0.2412, $BF_{10} =$

377   0.54; *sD*: p =0.3910, $BF_{10} = 0.41$; *fo*: p=0.1726, $BF_{10} = 0.59$). Again, the Bayes factor scores for

378   these three latter results are all less than 1, but greater than 0.33. As such, we cannot definitively

379   claim that there is no effect of attention on the *s*, *sD*, or *fo* representations. However, our data do

380   not provide any evidence of such an effect, unlike the strong effect we see for phonetic features.

381   This strong effect for phonetic features, but not the other representations, is clearly visible in

382   when visualizing the topographic distributions for the unique predictive power of each model

383   (Figure 3B).

384   We think the above approach of explicitly testing the unique predictive power of different speech

385   representations is a good way to explore the hierarchical processing of natural speech. However,

386   further insight could potentially be gleaned by investigating how the EEG is affected by attention

387   at different temporal delays relative to the speech stimulus. Indeed, in previous work we did this

388   by relating EEG to the speech envelope and finding attention effects at a specific range of time

389   lags from 195 – 230 ms, suggesting that cocktail party attention operates at relatively long

390   latencies, and, thus, beyond the earliest stages of acoustic processing (Power et al. 2012).

391   Similarly, we sought here to explore how the attention effects we have observed on our phonetic

392   features representation might vary across time lags. To do so, we first explored the possibility of

393   examining the weights of a TRF that captures the unique variance in the EEG due to the phonetic

394   features. To that end, we used cross-validation to fit a three-feature model based on *s*, *sD*, and *fo*

395   to best predict the EEG data. Then we subtracted this EEG prediction from the real EEG, and

396   again used cross validation to model the residual EEG using only the phonetic feature

397   representation. This model was able to predict significant variance in the residual EEG, and these

398   predictions were affected by attention, supporting our earlier results. However, there was no

399   clear effects of attention in the TRF weights themselves at any particular time lags. In fact, this

400   was not terribly surprising. In our previous studies on cocktail party attention, we have found

401   that attention effects on TRF weights themselves are generally much less robust and, thus,

402   require a large amount of data (Power et al. 2012), relative to approaches based on EEG

403   prediction or stimulus reconstruction using those TRFs (O'Sullivan et al. 2015).

404   With this in mind, we tried a second approach. Specifically, we repeated the approach of

405   predicting EEG using phonetic features *f*, while controlling for *s*, *sD*, and *fo*. However, this time,

406   rather than using a 0–300 ms interval of stimulus-response time lags, we ran the analysis

407   separately for different, non-overlapping 50 ms windows from 0–300 ms. Figure 3C shows that

408   attention has a significant effect on the phonetic feature-based predictions across a broad range

409   of time lags, starting from around 50 ms. However, the strongest effects of attention are visible

410   from 150-250 ms, with weaker effects from 50-100, 100-150, and from 200-250 ms, and no

411   attention effects at the shortest latencies of 0-50 ms (two-tailed Wilcoxon signed rank test; FDR

412   corrected; 0-50 ms: $p = 0.1726$, $BF_{10} = 0.5468$; 50-100 ms: $p = 0.0166$, $BF_{10} = 3.8610$; 100-150

413   ms: $p = 0.0017$, $BF_{10} = 11.0430$; 150-200 ms: $p = 6.1035e-4$, $BF_{10} = 66.8325$; 200-250 ms: $p =$

414   0.0031, $BF_{10} = 32.4147$; 250-300 ms: $p = 0.0203$, $BF_{10} = 3.6121$). This finding of strongest

415   attention effects at time lags in the range $150 – 250$ ms is largely consistent with earlier results

416   that identified strong cocktail party attention effects at temporal latencies beyond 100 ms with

417   limited effects at earlier latencies (Power et al. 2012; Puvvada and Simon 2017). One

418   complicating factor here, however, is that exploring EEG predictions using narrow ranges of

419   time delays is not necessarily as highly temporally resolved as it appears to be. This is because of

420   the autocorrelational structure of the speech stimuli and of the EEG. The stimulus-data

421   relationship at, say, 90 ms, can be quite similar to that at 110 ms. So the results need to be

422   interpreted with some sensitivity to the likelihood of smearing in time.

423   **[Fig 3 about here]**

424   Figure 3 | (A) Unique predictive power for the acoustic and phonetic feature spaces under two

425   attentional states. The features are as labelled on the horizontal axis. Statistical testing was

426   carried out to identify attentional effects (two-tailed Wilcoxon signed rank; *p<0.05, FDR

427   corrected). (B) Topographic distribution of partial correlations, averaged across all subjects.

428   (C) Unique predictive power for the phonetic features representation (*f*) in 50 ms windows from

429   0 – 300 ms, under two attentional states (two-tailed Wilcoxon signed rank test; *p<0.05,

430   **p<0.005, FDR corrected).

431   <u>Phonetic feature categories contribute predictive power beyond differentiating vowels and</u>

432   <u>consonants</u>

433   Di Liberto et al., (2015) found a high degree of discriminability between EEG responses to

434   vowels and consonants. The phonetic feature representation used in the above analysis

435   information on the manner, place, and voicing of consonants, as well as the articulatory position

436   of vowels. We wondered how our results would be affected if – instead of using a 19-

437   dimensional vector – the phonemes were simply marked as vowels or consonants. We repeated

438   the partial correlation analysis described above and found that there was a significant decrease in

439   prediction accuracy when the information on specific articulatory features was left out in the case

440   of attended speech, but not for unattended speech (two-tailed Wilcoxon signed-rank test,

441   p=0.011, z=2.542). Additionally, there was no longer a significant difference between the

442   prediction accuracies of attended and unattended stimuli (att_vc vs. unatt_vc; Wilcoxon signed-

443   rank test; p = 0.4326; z = 0.7847)

444   **[Fig 4 about here]**

445   Figure 4 | A significant reduction in unique predictive power was observed for the attended

446   condition when the partial correlation analysis was repeated with phonemes only categorised as

447   vowels or consonants (vc), as opposed to their underlying articulatory features (f) (two-tailed

448   Wilcoxon signed-rank test, *p=0.011, z=2.542).

449

450   **Discussion**

451   There has been longstanding debate on how selective attention affects the processing of speech.

452   Here, we set out to investigate attentional modulation at the pre-lexical level. In particular, based

453   on earlier work (Di Liberto et al., 2015; Daube et al., 2019), we considered how two acoustic and

454   two phonetic feature representations of attended and unattended speech were reflected in EEG

455   responses to that speech. We found that, for attended speech, including a 19-dimensional

456   phonetic features representation improved the prediction of the EEG responses beyond that

457   obtained when only using acoustic features. This was not true for unattended speech.

458   Furthermore, we found that the unique predictive power of the phonetic feature representation

459   was enhanced for attended versus unattended speech. This was not true for any other feature. We

460   contend that these findings make two important contributions to the literature. First, they

461   contribute to the debate around pre-lexical speech processing in cortex. And second, they

462   contribute to the longstanding debate on how selective attention affects the processing of speech.

463   In terms of prelexical speech processing, our study shows that a phonetic feature representation

464   has unique predictive power when it comes to modelling responses to attended speech (Fig 2 &

465   Table 1). This suggests that processing attended, but not unattended speech, may involve a

466   mapping from an acoustic to a categorical phonemic representation. This is a controversial idea.

467   In particular, while linguistic theories and prominent speech processing models posit a mapping

468   to phonetic features and/or phonemes as an intermediate stage between low-level acoustics and

469   words (McClelland and Elman 1986; A. M. Liberman et al. 1967; Hickok and Poeppel 2007),

470   this viewpoint is not unanimous. Some researchers question the necessity and existence of a

471   mental representation at this level for comprehending speech meaning (Lotto and Holt, 2000).

472   Indeed, there have been theories advocating for alternatives, including a different intermediate

473   representational unit (e.g. syllables; Massaro 1974), or direct matching to the lexicon based on

474   the comparison of acoustic representations (Goldinger 1998; Klatt 1989). In support of an

475   acoustic-phonemic stage of processing, there has been evidence from behavioral and EEG

476   studies that humans perceive phoneme categories. Liberman et al. (1957) found that

477   discrimination within phoneme categorical boundaries is poorer than across them. Eimas et al.

478   (1971) looked at infants' ability to discriminate syllables that differed in a voicing (phonetic)

479   feature and found that infants also perceived categorical phonetic features. Additionally,

17

480   Näätänen et al. (2001) found that the mismatch negativity (MMN), an event-related potential

481   component that reflects discriminable change in some repetitive aspect of the ongoing auditory

482   stimulation, is observed when the deviant stimulus is a phoneme prototype of a subject's native

483   language relative to when it is a non-prototype. However, it has been argued that these studies

484   used simple isolated units, and therefore their results could be due to the type of task and not

485   reflective of everyday listening.

486   More recent neuroimaging studies have shown that a phonetic feature representation of speech

487   can predict neural activity during listening to natural, continuous speech. Using high-density

488   cortical surface electrodes, Mesgarani et al. (2014) found that the superior temporal gyrus (STG)

489   selectively responds to phonetic features (their Figure 2-4). Di Liberto et al., (2015) and

490   Khalighinejad et al., (2017) found evidence that non-invasive EEG activity reflects the

491   categorization of speech into phonetic features. However, a recent paper challenged these

492   findings and argued that neural responses to speech could be well explained by a model that is

493   based entirely on acoustic features (Daube et al., 2019). Namely, they found that acoustic onsets

494   made very similar predictions to the benchmark phonetic features, and that acoustic onsets

495   explained parts of the neural response that phonetic features could not. Here, we show evidence

496   to support the encoding of a rich array of phonetic features (certainly more than simply vowel

497   and consonant categorization; Fig. 4) for attended speech. It is possible that we were able to find

498   unique predictive power for these phonetic features because our study involved stimuli from two

499   different talkers (a male and a female) while the stimuli used in Daube et al. (2019) were spoken

500   by a single talker. As mentioned earlier, there is redundancy between speech acoustics and a

501   phonetic feature representation because each phonetic feature has a characteristic spectro-

502   temporal acoustic profile. This overlap can make it difficult to isolate the precise contribution of

503   each predictor. Indeed, if there were no variation in spectro-temporal acoustics across each

504   utterance of the same phoneme, the phonetic feature and spectrogram representations would be,

505   effectively, identical. However, the more variation there is between utterances of the same

506   phoneme (which will happen when including both male and female speakers), the greater the

507   difference should be between the contribution of neurons that care about those variations (e.g.,

508   low-level "acoustic" neurons) and the contribution of neurons whose responses are invariant to

509   the acoustic details and only dependent on phoneme category. As such, it may be that, when

510   using stimuli with high intra-phoneme acoustic variation, including both acoustic and phonetic

18

511 feature representations is of most benefit for predicting neural activity. Future work will
512 investigate this idea further.

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

532 As a final comment on this issue, we acknowledge that there are also other acoustic
533 transformations of the auditory stimulus that have not been considered beyond the spectrogram
534 and its derivative, as well as other theorized intermediate representations apart from
535 phonemes/phonetic features that have not been tested. It is likely that some of these feature sets
536 may overlap with our measure of phonetic features. As such, we cannot definitively prove the
537 encoding of this measure in cortex. However, one reason the idea of mapping to words via an
538 acoustic-phonetic stage has prevailed is due to its computational efficiency (Scharenborg et al.
539 2005). It is less obvious how the detection of acoustic onsets and other acoustic representations

540   would lead to lexical access without an intermediate stage – and in particular, how such a model

541   would robustly generalize to new words and new speakers.

542   As to the cocktail party attention debate, our results suggest that attention differentially

543   modulates cortical processing of acoustic and phonetic information. We found that the *unique*

544   predictive power of phonetic features was modulated by attention, while the other feature spaces

545   were not. This result is in line with a wealth of evidence for stronger attention effects at higher

546   levels of the speech processing hierarchy. For example, there is no doubt that both attended and

547   unattended speech will influence activity in the cochlea and the auditory nerve. But by the time

548   one reaches auditory association areas like STG, only the attended speaker is represented

549   (Mesgarani and Chang 2012; Zion Golumbic et al. 2013). Indeed, an important recent study (J.

550   O'Sullivan et al. 2019) used simultaneous recordings in both primary auditory cortex (Heschl's

551   gyrus; HG) and STG, to show that both attended and unattended speech are robustly represented

552   in primary areas, with STG selectively representing the attended speaker. This lines up very well

553   with our results. In particular, it is well established that spectro-temporal models of sounds can

554   do a good job of explaining neural responses in primary auditory areas, but not in nonprimary

555   auditory cortex (Norman-Haignere and McDermott 2018). Thus, we might expect our isolated

556   acoustic speech processing indices to mostly reflect activity from primary auditory areas, which

557   have been shown to be relatively unaffected by attention (J. O'Sullivan et al. 2019). Meanwhile

558   STG is known to be selective for phonetic feature representations (Mesgarani et al. 2014) and

559   activity in STG is strongly affected by attention (J. O'Sullivan et al. 2019). Thus, we might

560   expect our isolated measures of phonetic feature processing to be strongly influenced by activity

561   from STG, and thus, significantly affected by attention. This is the pattern of results that we have

562   observed.

563   Additional support for our findings as evidence for a higher-level locus of attentional selection

564   comes from research examining the latency of attention effects on EEG and MEG responses to

565   speech. This approach has commonly focused on a rather general measure of speech processing

566   that derives from indexing how the neural data track the speech envelope. Directly determining

567   how much this speech processing is driven by acoustic versus speech-specific processing is not

568   clear, although there is undoubtedly a very substantial acoustic contribution (Lalor et al. 2009;

569   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 (e.g., 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 less than 1, they were not less than 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, Ehud, Srikantan Nagarajan, Merav Ahissar, Athanassios Protopapas, Henry Mahncke, and Michael M. Merzenich. 2001. "Speech Comprehension Is Correlated with Temporal Response Patterns Recorded from Auditory Cortex." *Proceedings of the National Academy of Sciences of the United States of America* 98 (23): 13367–72. https://doi.org/10.1073/pnas.201400998.

602 AJE, Kell, Yamins DLK, Shook EN, Norman-Haignere SV, and McDermott JH. 2018. "A Task-
603   Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain
604   Responses, and Reveals a Cortical Processing Hierarchy." *Neuron* 98 (3): 630-644.e16.
605   https://doi.org/10.1016/J.NEURON.2018.03.044.

606 Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF
607   database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing
608   to Audio and Acoustics (Cat. No.01TH8575)*, 99–102.

609 Broadbent, D. (1958). Perception and Communication. Pergamon Press, London, UK

610 Brodbeck, Christian, L. Elliot Hong, and Jonathan Z. Simon. 2018. "Rapid Transformation from
611   Auditory to Linguistic Representations of Continuous Speech." *Current Biology* 28 (24):
612   3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042.

613 Broderick, Michael P., Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and
614   Edmund C. Lalor. 2018. "Electrophysiological Correlates of Semantic Dissimilarity Reflect
615   the Comprehension of Natural, Narrative Speech." *Current Biology* 28 (5): 803-809.e3.
616   https://doi.org/10.1016/j.cub.2018.01.080.

617 C, Tang, Hamilton LS, and Chang EF. 2017. "Intonational Speech Prosody Encoding in the
618   Human Auditory Cortex." *Science (New York, N.Y.)* 357 (6353): 797–801.
619   https://doi.org/10.1126/SCIENCE.AAM8577.

620 Chang, Edward F, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and
621   Robert T Knight. 2010. "Categorical Speech Representation in Human Superior Temporal
622   Gyrus." *Nature Neuroscience* 13 (11): 1428–32. https://doi.org/10.1038/nn.2641.

623 Cherry, E Colin. 1953. "Some Experiments on the Recognition of Speech, with One and with
624   Two Ears." *Citation: The Journal of the Acoustical Society of America* 25: 975.
625   https://doi.org/10.1121/1.1907229.

626 Chomsky, N, and M Halle. 1968. "1968: The Sound Pattern of English. New York: Harper and
627   Row."

628 Crosse, Michael J., Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. 2016. "The
629   Multivariate Temporal Response Function (MTRF) Toolbox: A MATLAB Toolbox for
630   Relating Neural Signals to Continuous Stimuli." *Frontiers in Human Neuroscience* 10
631   (November): 604. https://doi.org/10.3389/fnhum.2016.00604.

632 Daube, Christoph, Robin A A Ince, and Joachim Gross. 2019. "Simple Acoustic Features Can
633   Explain Phoneme-Based Predictions of Cortical Responses to Speech Article Simple
634   Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to
635   Speech." *Current Biology* 29. https://doi.org/10.1016/j.cub.2019.04.067.

636 Davis, Matthew H., and Ingrid S. Johnsrude. 2003. "Hierarchical Processing in Spoken
637   Language Comprehension." *Journal of Neuroscience* 23 (8): 3423–31.
638   https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003.

639 Delorme, Arnaud, and Scott Makeig. 2004. "EEGLAB: An Open Source Toolbox for Analysis
640   of Single-Trial EEG Dynamics Including Independent Component Analysis." *Journal of
641   Neuroscience Methods* 134 (1): 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

642 Deutsch, J. A., and D. Deutsch. 1963. "Attention: Some Theoretical Considerations."
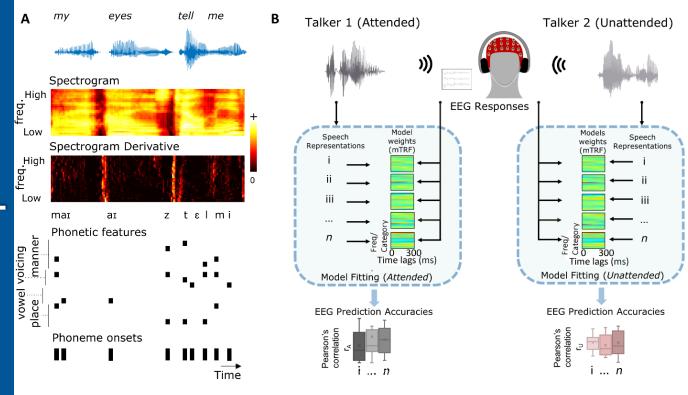643   *Psychological Review* 70 (1): 80–90. https://doi.org/10.1037/H0039515.

644 DeWitt, Iain, and Josef P. Rauschecker. 2012. "Phoneme and Word Recognition in the Auditory
645   Ventral Stream." *Proceedings of the National Academy of Sciences of the United States of*
646   *America* 109 (8): 2709. https://doi.org/10.1073/pnas.1113427109.

647 Di Liberto, Giovanni M., James A. O'Sullivan, and Edmund C. Lalor. 2015. "Low-Frequency
648   Cortical Entrainment to Speech Reflects Phoneme-Level Processing." *Current Biology* 25
649   (19): 2457–65. https://doi.org/10.1016/J.CUB.2015.08.030.

650 Ding, Nai, and Jonathan Z. Simon. 2012. "Neural Coding of Continuous Speech in Auditory
651   Cortex during Monaural and Dichotic Listening." *Journal of Neurophysiology* 107 (1): 78–
652   89. https://doi.org/10.1152/jn.00297.2011.

653 EC, Lalor, Power AJ, Reilly RB, and Foxe JJ. 2009. "Resolving Precise Temporal Processing
654   Properties of the Auditory System Using Continuous Stimuli." *Journal of Neurophysiology*
655   102 (1): 349–59. https://doi.org/10.1152/JN.90896.2008.

656 ES, Teoh, and Lalor EC. 2019. "EEG Decoding of the Target Speaker in a Cocktail Party
657   Scenario: Considerations Regarding Dynamic Switching of Talker Location." *Journal of*
658   *Neural Engineering* 16 (3). https://doi.org/10.1088/1741-2552/AB0CF1.

659 ES, Teoh, Cappelloni MS, and Lalor EC. 2019. "Prosodic Pitch Processing Is Represented in
660   Delta-Band EEG and Is Dissociable from the Cortical Tracking of Other Acoustic and
661   Phonetic Features." *The European Journal of Neuroscience* 50 (11): 3831–42.
662   https://doi.org/10.1111/EJN.14510.

663 Fisher, Ronald Aylmer, Sir, 1890-1962. 1924. "035: The Distribution of the Partial Correlation
664   Coefficient." https://digital.library.adelaide.edu.au/dspace/handle/2440/15182.

665 Gwilliams, Laura, Jean-Remi King, Alec Marantz, and David Poeppel. 2020. "Neural Dynamics
666   of Phoneme Sequencing in Real Speech Jointly Encode Order and Invariant Content."
667   *BioRxiv*, April, 2020.04.04.025684. https://doi.org/10.1101/2020.04.04.025684.

668 Hickok, Gregory, and David Poeppel. 2007. "The Cortical Organization of Speech Processing."
669   *Nature Reviews Neuroscience* 8 (5): 393–402. https://doi.org/10.1038/nrn2113.

670 Holdgraf, Christopher R., Jochem W. Rieger, Cristiano Micheli, Stephanie Martin, Robert T.
671   Knight, and Frederic E. Theunissen. 2017. "Encoding and Decoding Models in Cognitive
672   Electrophysiology." *Frontiers in Systems Neuroscience* 11 (September): 61.
673   https://doi.org/10.3389/FNSYS.2017.00061.

674 Johnston, William A., and Jolynn Wilson. 1980. "Perceptual Processing of Nontargets in an
675   Attention Task." *Memory & Cognition 1980 8:4* 8 (4): 372–77.
676   https://doi.org/10.3758/BF03198277.

677 JR, Binder, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, and Possing ET.
678   2000. "Human Temporal Lobe Activation by Speech and Nonspeech Sounds." *Cerebral*
679   *Cortex (New York, N.Y. : 1991)* 10 (5): 512–28. https://doi.org/10.1093/CERCOR/10.5.512.

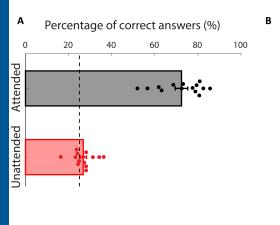680 K, Okada, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, and Hickok G.

681   2010. "Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic
682   Invariance in the Response to Intelligible Speech." *Cerebral Cortex (New York, N.Y. : 1991)*
683   20 (10): 2486–95. https://doi.org/10.1093/CERCOR/BHP318.

684 Khalighinejad, Bahar, Guilherme Cruzatto da Silva, and Nima Mesgarani. 2017. "Dynamic
685   Encoding of Acoustic Features in Neural Responses to Continuous Speech." *Journal of*
686   *Neuroscience* 37 (8): 2176–85. https://doi.org/10.1523/JNEUROSCI.2383-16.2017.

687 Lalor, Edmund C., and John J. Foxe. 2010. "Neural Responses to Uninterrupted Natural Speech
688   Can Be Extracted with Precise Temporal Resolution." *European Journal of Neuroscience*
689   31 (1): 189–93. https://doi.org/10.1111/j.1460-9568.2009.07055.x.

690 Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. 1967. "Perception
691   of the Speech Code." *Psychological Review* 74 (6): 431–61.
692   https://doi.org/10.1037/H0020279.

693 Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith.
694   1957. "The Discrimination of Speech Sounds within and across Phoneme Boundaries."
695   *Journal of Experimental Psychology* 54 (5): 358–68. https://doi.org/10.1037/H0044417.

696 Liberto, Giovanni M. Di, Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné. 2019.
697   "Low-Frequency Cortical Responses to Natural Speech Reflect Probabilistic Phonotactics."
698   *NeuroImage* 196 (August): 237–47. https://doi.org/10.1016/J.NEUROIMAGE.2019.04.037.

699 Lotto, A. J., & Holt, L. (2000). The Illusion of the Phoneme. In: Billing, S.J., Boyle, J.P., &
700   Griffith, A.M. (eds), *Chicago Linguistic Society*, 35, 191-204.

701 Massaro, Dominic W. 1974. "Perceptual Units in Speech Recognition." *Journal of Experimental*
702   *Psychology* 102 (2): 199–208. https://doi.org/10.1037/H0035854.

703 McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan
704   Sonderegger. 2017. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using
705   Kaldi." *Proceedings of the Annual Conference of the International Speech Communication*
706   *Association, INTERSPEECH* 2017-August: 498–502.
707   https://doi.org/10.21437/INTERSPEECH.2017-1386.

708 McClelland, James L., and Jeffrey L. Elman. 1986. "The TRACE Model of Speech Perception."
709   *Cognitive Psychology* 18 (1): 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.

710 Mesgarani, Nima, and Edward F. Chang. 2012. "Selective Cortical Representation of Attended
711   Speaker in Multi-Talker Speech Perception." *Nature*. Nature Publishing Group.
712   https://doi.org/10.1038/nature11020.

713 Mesgarani, Nima, Connie Cheung, Keith Johnson, and Edward F. Chang. 2014. "Phonetic
714   Feature Encoding in Human Superior Temporal Gyrus." *Science* 343 (6174): 1006–10.
715   https://doi.org/10.1126/science.1245994.

716 MF, Howard, and Poeppel D. 2010. "Discrimination of Speech Stimuli Based on Neuronal
717   Response Phase Patterns Depends on Acoustics but Not Comprehension." *Journal of*
718   *Neurophysiology* 104 (5): 2500–2511. https://doi.org/10.1152/JN.00251.2010.

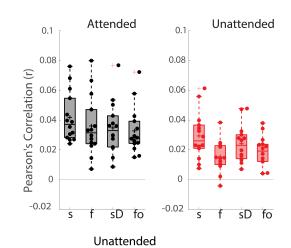719 Moray, Neville. 2018. "Attention in Dichotic Listening: Affective Cues and the Influence of

720   Instructions:" *Https://Doi.Org/10.1080/17470215908416289* 11 (1): 56–60.
721   https://doi.org/10.1080/17470215908416289.

722   MR, Schädler, and Kollmeier B. 2015. "Separable Spectro-Temporal Gabor Filter Bank
723   Features: Reducing the Complexity of Robust Features for Automatic Speech Recognition."
724   *The Journal of the Acoustical Society of America* 137 (4): 2047–59.
725   https://doi.org/10.1121/1.4916618.

726   Norman-Haignere, Sam V., and Josh H. McDermott. 2018. "Neural Responses to Natural and
727   Model-Matched Stimuli Reveal Distinct Computations in Primary and Nonprimary
728   Auditory Cortex." *PLOS Biology* 16 (12): e2005127.
729   https://doi.org/10.1371/JOURNAL.PBIO.2005127.

730   O'Sullivan, James A., Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe,
731   Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C.
732   Lalor. 2015. "Attentional Selection in a Cocktail Party Environment Can Be Decoded from
733   Single-Trial EEG." *Cerebral Cortex* 25 (7): 1697–1706.
734   https://doi.org/10.1093/cercor/bht355.

735   O'Sullivan, James, Jose Herrero, Elliot Smith, Catherine Schevon, Guy M. McKhann, Sameer A.
736   Sheth, Ashesh D. Mehta, and Nima Mesgarani. 2019. "Hierarchical Encoding of Attended
737   Auditory Objects in Multi-Talker Speech Perception." *Neuron* 104 (6): 1195-1209.e3.
738   https://doi.org/10.1016/j.neuron.2019.09.007.

739   PD, Eimas, Siqueland ER, Jusczyk P, and Vigorito J. 1971. "Speech Perception in Infants."
740   *Science (New York, N.Y.)* 171 (3968): 303–6.
741   https://doi.org/10.1126/SCIENCE.171.3968.303.

742   Peelle, Jonathan E, Ingrid Johnsrude, and Matthew H Davis. 2010. "Hierarchical Processing for
743   Speech in Human Auditory Cortex and Beyond." *Frontiers in Human Neuroscience* 4
744   (June): 51. https://doi.org/10.3389/FNHUM.2010.00051.

745   Power, Alan J., John J. Foxe, Emma-Jane Forde, Richard B. Reilly, and Edmund C. Lalor. 2012.
746   "At What Time Is the Cocktail Party? A Late Locus of Selective Attention to Natural
747   Speech." *European Journal of Neuroscience* 35 (9): 1497–1503.
748   https://doi.org/10.1111/j.1460-9568.2012.08060.x.

749   Power, Alan J., Edmund C. Lalor, and Richard B. Reilly. 2011. "Endogenous Auditory Spatial
750   Attention Modulates Obligatory Sensory Activity in Auditory Cortex." *Cerebral Cortex* 21
751   (6): 1223–30. https://doi.org/10.1093/CERCOR/BHQ233.

752   Prinsloo, Kevin D., and Edmund C. Lalor. 2020. "General Auditory and Speech-Specific
753   Contributions to Cortical Envelope Tracking Revealed Using Auditory Chimeras." *BioRxiv*,
754   October, 2020.10.21.348557. https://doi.org/10.1101/2020.10.21.348557.

755   Puvvada, Krishna C., and Jonathan Z. Simon. 2017. "Cortical Representations of Speech in a
756   Multitalker Auditory Scene." *Journal of Neuroscience* 37 (38): 9189–96.
757   https://doi.org/10.1523/JNEUROSCI.0938-17.2017.

758   R, Näätänen. 2001. "The Perception of Speech Sounds by the Human Brain as Reflected by the
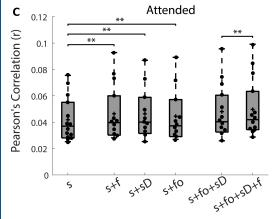759   Mismatch Negativity (MMN) and Its Magnetic Equivalent (MMNm)." *Psychophysiology*

760   38 (1): 1–21. https://doi.org/10.1017/S0048577201000208.

761 Rauschecker, Josef P., and Sophie K. Scott. 2009. "Maps and Streams in the Auditory Cortex:
762   Nonhuman Primates Illuminate Human Speech Processing." *Nature Neuroscience*.
763   https://doi.org/10.1038/nn.2331.

764 Reddy, Sravana, and James N. Stanford. 2015. "Toward Completely Automated Vowel
765   Extraction: Introducing DARLA." *Linguistics Vanguard* 1 (1): 15–28.
766   https://doi.org/10.1515/LINGVAN-2015-0002.

767 Klatt, D. H. (1989). Review of selected models of speech perception. In *Lexical representation*
768   *and process* (pp. 169–226). The MIT Press.

769 Richard Morey, Maintainer D. 2018. "Package 'BayesFactor' Title Computation of Bayes
770   Factors for Common Designs."

771 Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary
772   Prichard, and Jiahong Yuan. 2014. "FAVE (Forced Alignment and Vowel Extraction) Suite
773   Version 1.1.3," May. https://doi.org/10.5281/ZENODO.9846.

774 Rouder, Jeffrey N., Richard D. Morey, Paul L. Speckman, and Jordan M. Province. 2012.
775   "Default Bayes Factors for ANOVA Designs." *Journal of Mathematical Psychology* 56 (5):
776   356–74. https://doi.org/10.1016/J.JMP.2012.08.001.

777 Scharenborg, Odette, Dennis Norris, Louis ten Bosch, and James M. McQueen. 2005. "How
778   Should a Speech Recognizer Work?" *Cognitive Science* 29 (6): 867–918.
779   https://doi.org/10.1207/S15516709COG0000_37.

780 SD, Goldinger. 1998. "Echoes of Echoes? An Episodic Theory of Lexical Access."
781   *Psychological Review* 105 (2): 251–79. https://doi.org/10.1037/0033-295X.105.2.251.

782 Treisman, A. M. 1964. "Verbal cues, language, and meaning in selective attention." *The*
783   *American Journal of Psychology* 77: 206–19. https://doi.org/10.2307/1420127.

784 Tune, Sarah, Mohsen Alavash, Lorenz Fiedler, and Jonas Obleser. 2020. "Neural Attention
785   Filters Do Not Predict Behavioral Success in a Large Cohort of Aging Listeners." *BioRxiv*,
786   May, 2020.05.20.105874. https://doi.org/10.1101/2020.05.20.105874.

787 Wagenmakers, Eric Jan, Ruud Wetzels, Denny Borsboom, and Han L.J. van der Maas. 2011.
788   "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi:
789   Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100 (3): 426–32.
790   https://doi.org/10.1037/A0022790.

791 Y. Benjamini, Y. Hochberg 1995. "Controlling the False Discovery Rate : A Practical and
792   Powerful Approach to Multiple Testing". J. R. Stat. Soc. Series B Stat.
793   Methodol., 57 (1995), pp. 289-300

794 Zion Golumbic, Elana M., Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A. Schevon, Guy
795   M. McKhann, Robert R. Goodman, et al. 2013. "Mechanisms Underlying Selective
796   Neuronal Tracking of Attended Speech at a 'Cocktail Party.'" *Neuron* 77 (5): 980–91.
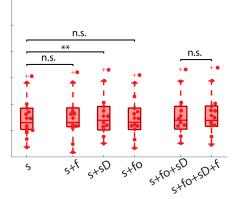797   https://doi.org/10.1016/j.neuron.2012.12.037.

798

799

800 **Acknowledgments**

805

A

my   *eyes*   *tell* *me*

Spectrogram

High

freq.

Low

+

0

Spectrogram Derivative

High

freq.

Low

maɪ   aɪ   z t ɛ l m i

Phonetic features

manner

voicing

vowel

place

Phoneme onsets

Time

B

Talker 1 (Attended)   Talker 2 (Unattended)

EEG Responses

Speech   Model   Models   Speech
Representations   weights   weights   Representations
                        (mTRF)   (mTRF)

i   i

ii   ii

iii   iii

...   ...

n   n

Freq/   Freq/
Category   Category

0   300   0   300
Time lags (ms)   Time lags (ms)

Model Fitting (*Attended*)   Model Fitting (*Unattended*)

EEG Prediction Accuracies   EEG Prediction Accuracies

Pearson's   Pearson's
correlation   correlation
r_A r_U

i ... *n*   i ... *n*

A Percentage of correct answers (%)

B

Attended

Unattended

Pearson's Correlation (r)

s   f   sD   fo

s   f   sD   fo

C

Attended

Pearson's Correlation (r)

s   s+f   s+sD   s+fo   s+fo+sD   s+fo+sD+f

Unattended

s   s+f   s+sD   s+fo   s+fo+sD   s+fo+sD+f

s   : spectrogram

sD : spectrogram derivative

f   : phonetic features

fo : phoneme onsets

A

B

C