Title

The effect of prior knowledge and intelligibility on the cortical entrainment response to speech

Journal

Journal of Neurophysiology, 118(6)

ISSN

0022-3077

Authors

Baltzell, Lucas S

Srinivasan, Ramesh

Richards, Virginia M

Publication Date

2017-12-01

DOI

10.1152/jn.00023.2017

J Neurophysiol 118: 3144–3151, 2017.
First published September 6, 2017; doi:10.1152/jn.00023.2017.

RESEARCH ARTICLE | *Sensory Processing*

The effect of prior knowledge and intelligibility on the cortical entrainment response to speech

Lucas S. Baltzell,[1] Ramesh Srinivasan,[1,2] and Virginia M. Richards[1]
[1]*Department of Cognitive Sciences, University of California, Irvine, California; and* [2]*Department of Biomedical Engineering, University of California, Irvine, California*

Submitted 10 January 2017; accepted in final form 1 September 2017

Baltzell LS, Srinivasan R, Richards VM. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *J Neurophysiol* 118: 3144–3151, 2017. First published September 6, 2017; doi:10.1152/jn.00023.2017.—It has been suggested that cortical entrainment plays an important role in speech perception by helping to parse the acoustic stimulus into discrete linguistic units. However, the question of whether the entrainment response to speech depends on the intelligibility of the stimulus remains open. Studies addressing this question of intelligibility have, for the most part, significantly distorted the acoustic properties of the stimulus to degrade the intelligibility of the speech stimulus, making it difficult to compare across "intelligible" and "unintelligible" conditions. To avoid these acoustic confounds, we used priming to manipulate the intelligibility of vocoded speech. We used EEG to measure the entrainment response to vocoded target sentences that are preceded by natural speech (nonvocoded) prime sentences that are either valid (match the target) or invalid (do not match the target). For unintelligible speech, valid primes have the effect of restoring intelligibility. We compared the effect of priming on the entrainment response for both 3-channel (unintelligible) and 16-channel (intelligible) speech. We observed a main effect of priming, suggesting that the entrainment response depends on prior knowledge, but not a main effect of vocoding (16 channels vs. 3 channels). Furthermore, we found no difference in the effect of priming on the entrainment response to 3-channel and 16-channel vocoded speech, suggesting that for vocoded speech, entrainment response does not depend on intelligibility.

NEW & NOTEWORTHY Neural oscillations have been implicated in the parsing of speech into discrete, hierarchically organized units. Our data suggest that these oscillations track the acoustic envelope rather than more abstract linguistic properties of the speech stimulus. Our data also suggest that prior experience with the stimulus allows these oscillations to better track the stimulus envelope.

attention; EEG; entrainment; intelligibility; prior knowledge

INTRODUCTION

Synchronous oscillations in the cortex arise naturally from both synaptic delays in local neural circuits and transmission delays in long-range neural connections (Nunez and Srinivasan 2006). For certain neural circuits, it has been demonstrated that oscillatory activity reflects synchronized modulations of neuronal excitability (Buzsáki and Draguhn 2004). Based in large part on these findings and on findings that attention can lead to increased neural synchrony (e.g., Fries et al. 2001), it has been suggested that oscillations enable coordination at different timescales across and within cortical networks, and in this way govern perceptual and cognitive processes (e.g., Fries 2005). Lending support to this claim, recent studies have shown that cortical oscillations can hierarchically couple such that the phase of lower frequency oscillations can modulate the power at higher frequency oscillation, and that this coupling is task specific (Canolty et al. 2006; Lakatos et al. 2005, 2008). Although this literature has generated some claims that remain controversial, it has also proposed a set of neural mechanisms through which oscillations can govern perceptual processes.

It has been proposed that phase-resetting cortical oscillations support speech perception by aligning fluctuations of neuronal excitability to periodic changes in the speech stimulus, a process that has been labeled cortical entrainment (Ding and Simon 2014). Cortical entrainment is thought to support the parsing of acoustic input into discrete linguistic units such as phrases, syllables, and phonemes by aligning hierarchically coupled oscillations at multiple timescales to the natural fluctuations of the speech stimulus (e.g., Ghitza, 2011; Giraud and Poeppel 2012). Supporting this claim, Ding et al. (2016) identified a neural response that entrains to periodic fluctuations in high-level linguistic units (including verb phrases and sentences) in the absence of acoustic cues, distinct from a neural response that entrains to the acoustic envelope regardless of linguistic content. Although this study elegantly differentiated between entrainment to linguistic and acoustic cues, the speech stimuli used to make this distinction were artificially periodic. Specifically, the stimuli were single words presented at a rate of 4 Hz, with phrases constituting 2-word units (2-Hz repetition rate) and sentences constituting 4-word units (1-Hz repetition rate). Modulation rates in natural speech, however, are less regular and less predictable, as are fluctuations in linguistic units such as phrases and sentences, making it difficult to generalize these results to natural speech.

For natural speech, it has proved difficult to demonstrate entrainment to linguistic features. A number of studies have failed to show an effect of intelligibility on entrainment, suggesting that entrainment may not reflect the tracking of linguistic features. For instance, Howard and Poeppel (2010) showed that entrainment to natural speech is no greater than entrainment to time-reversed speech (see also Zoefel and

Address for reprint requests and other correspondence: L. S. Baltzell, Social Science Lab (SSL) 184, Univ. of California, Irvine, CA 92697 (e-mail: baltzell@uci.edu).

VanRullen 2016; but see Hertrich et al. 2013). Furthermore, Doelling et al. (2014) showed that entrainment to click trains, which carry no linguistic information but contain maximally abrupt acoustic changes, was significantly higher than entrainment to natural speech, suggesting that entrainment may reflect the abruptness of acoustic transitions rather than linguistic content.

Other studies have shown that degrading the intelligibility of speech stimulus leads to a decrease in the strength of entrainment to that stimulus. However, we know that entrainment to acoustic changes can occur in the absence of linguistic information (e.g., Doelling et al. 2014), so when manipulations in intelligibility also fundamentally alter the acoustics of the speech stimulus, it is difficult to conclude that the effects of this manipulation on entrainment arise from the degradation of linguistic cues. For instance, Peelle et al. (2013) showed that entrainment to intelligible vocoded speech (16 channels) is significantly greater than entrainment to unintelligible vocoded speech (1 channel). In this case, it is unclear whether entrainment is driven by intelligibility (and therefore reflects a linguistic contribution) because vocoding degrades the acoustic input to the auditory system, and this degradation is more severe for speech vocoded with lower numbers of channels (Shannon et al. 1995). It is therefore possible that the decrease in entrainment response from 16-channel to 1-channel speech reflects a degradation in acoustic input rather than a degradation of intelligibility. Such a concern can also be raised when interpreting the results of Ahissar et al. (2001), who showed that entrainment to time-compressed speech was reduced relative to natural speech. Again, time compression is a manipulation that distorts the acoustic input to the auditory system, making it difficult to determine whether the reduction in entrainment reflects degraded linguistic information or degraded acoustic information.

In an attempt to avoid these acoustic confounds, Millman et al. (2015) used the same two vocoded sentences before and after perceptual learning to measure the effect of intelligibility on the entrainment response. They used a 3-channel vocoder to ensure that the sentences were unintelligible before training. Before training, these vocoded sentences were presented over multiple trials and the MEG response was recorded. During training, one of these vocoded sentences was then paired with the nonvocoded original so that the meaning of the sentence could be understood despite the vocoding. After training, these vocoded stimuli were once again presented to the listener while the MEG response was recorded such that one vocoded sentence was now intelligible while the other remained unintelligible. They found no difference in the strength of entrainment to the posttraining intelligible sentence and the entrainment response to the posttraining unintelligible sentence, suggesting that entrainment is driven by speech acoustic rather than linguistic cues.

We considered it possible, however, that the failure of Millman et al. (2015) to find an effect of intelligibility on the entrainment response might be due to overlearning through the use of limited speech materials. We know that the neural response to verbal materials can decrease after overlearning (Thompson and Thompson 1965), and we thought it possible that mechanisms involved with speech comprehension may be less active after overlearning. We also considered that a lack of a demanding behavioral task may have limited the amount of attention listeners deployed in encoding the stimuli, which may also have contributed to the lack of effect. In the current study, we attempted to remove these potential confounds while also employing a design that allowed us to contrast effects of prior knowledge and effects of intelligibility on the entrainment response.

Following Millman et al. (2015), we used tone vocoding to degrade the intelligibility of spoken sentences. Sentences were drawn from the TIMIT database (Garofolo et al. 1993), which contains sentences from 630 different speakers of American English. Vocoding has been widely used as a procedure for manipulating intelligibility, and previous research has shown that whereas 16-channel vocoded speech is perfectly intelligible, 3-channel vocoded speech is largely unintelligible (Loizou et al. 1999). To differentially study the effect of prior knowledge and intelligibility, both 16-channel and 3-channel vocoded sentences were used.

To manipulate the intelligibility of 3-channel vocoded sentences without altering their acoustic properties, we primed the degraded (vocoded) sentences with nondegraded (natural speech) sentences. Although text priming has been used in the literature (e.g., Sohoglu et al. 2012), we chose not to use text priming because of the neuroanatomical differences between reading and spoken language comprehension, particularly at early stages of acoustic/phonetic processing (Buchweitz et al. 2009; Cohen et al. 2002; Price 2012), and the fact that we did not want cortical networks not typically involved with speech perception to influence our neural recordings. When a vocoded sentence is preceded by a nonvocoded copy of itself (valid condition), the intelligibility of the vocoded sentence is largely restored. Conversely, when the vocoded sentence is preceded by a nonvocoded sentence that is unrelated to the vocoded sentence (invalid condition), the intelligibility of the vocoded sentence is unaffected (Remez et al. 1981). We tested both of these conditions (valid and invalid).

Manipulating intelligibility using valid and invalid primes does not alter the acoustics of the target vocoded sentences. It does, however, alter the a priori expectations brought to bear on the vocoded sentences. To capture the effects of this prior knowledge, we use 16-channel vocoded speech under the same priming manipulation. Because 16-channel vocoded speech is intelligible, any difference between valid and invalid conditions can be attributed to this prior knowledge. This 2 (valid vs. invalid) $\times$ 2 (3-channel vs. 16-channel) design allows us to isolate the effects of intelligibility and prior knowledge in the absence of acoustic confounds, enabling us to measure the degree to which the entrainment response to ongoing speech tracks acoustic vs linguistic features of the stimulus. If the entrainment response depends on intelligibility, we expect a significant interaction between prime validity and number of vocoded channels such that a valid prime leads to a greater increase in the entrainment response to 3-channel vocoded speech than to 16-channel vocoded speech.

Finally, to motivate listeners to attend to the vocoded stimuli, after each vocoded sentence, a short clip of vocoded material (probe) was presented and listeners were asked to indicate whether or not that clip came from the vocoded sentence just heard on that trial. In an attempt to control for variation in task difficulty across conditions, an adaptive tracking procedure was used with separate tracks for each condition. Following a 2-down/1-up tracking procedure, the duration of

the probe was adaptively varied according to the listener's response, converging on the duration necessary to achieve 71% correct (Levitt 1971).

METHODS

Participants. All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Fourteen young adults (6 women; age 23–28 yr) participated in the study, although one (man) was excluded due to excessive EEG artifacts. Participants were not screened for handedness.

Stimuli. Speech materials were drawn from the TIMIT database (Garofolo et al. 1993), which contains sentences from 630 different speakers of American English. Specifically, we selected all sentences between 3.5 and 5 s in duration, yielding a total of 1,247 sentences. Many of the TIMIT sentences contain brief silent periods at the start of the recording, and these silent periods were removed before testing for all sentences used in the experiment.

We used a tone-vocoder algorithm with logarithmic band spacing developed by Nie et al. (2005) to manipulate the intelligibility of these sentences. Vocoding is a four-step process that *1*) bandpass filters the input stimulus into individual frequency bands, *2*) extracts a low-pass filtered envelope from the output of each filter, *3*) modulates a tone carrier with the resulting envelope, and *4*) bandpass filters the resulting bands and sums them together. Both 3-channel and 16-channel vocoders were used in the experiment. In our implementation of this algorithm, we specified a 30-Hz low-pass filter cutoff.

Task. Each participant sat in a single-walled sound-attenuated booth, and stimuli were presented diotically at 70 dB SPL over Stax electrostatic headphones. Testing was conducted over two sessions, each of which contained four blocks. For each session, two blocks contained exclusively 16-channel vocoded trials and two blocks contained exclusively 3-channel vocoded trials (order was randomized). Before each block of EEG recording, participants were asked to rate the perceived intelligibility of target vocoded sentences in both the valid and invalid prime conditions on a scale from 1 to 6, where a 6 indicated that they understood all of the words in the sentence and 1 indicated that they understood none of the words. On valid trials (when the target and prime matched), participants were asked not to consider the intelligibility of the prime when making their intelligibility judgements. Because 16-channel vocoded speech is intelligible regardless of priming, in the interest of time, we only obtained a single intelligibility rating for each condition (valid vs. invalid prime) before 16-channel experimental blocks. Before 3-channel blocks, however, we obtained intelligibility ratings for 15 target sentences in each condition to provide a stable estimate. Intelligibility ratings were gathered to ensure that *1*) in the 3-channel vocoding condition, valid primes led to higher intelligibility ratings than invalid primes, and *2*) this effect was stable over blocks. Because of the high variability of individual talkers in our stimulus set, we expected minimal adaptation to the vocoded stimuli.

On each trial, target vocoded sentences were preceded by a natural speech prime sentence and followed by a probe clip of vocoded speech material (Fig. 1). On half the trials, these natural speech prime sentences matched the target vocoded sentence (valid prime), and on the other half, they did not match (invalid prime). Participants were asked to indicate whether or not the probe was drawn from the target vocoded sentence, and the duration of the probe was adaptively varied following a 2-down/1-up procedure, which converges on the duration necessary to achieve 71% correct (Levitt 1971). We used a ratio step size of 1.2 such that increases in probe length were in steps of 1.2 times the previous length, and decreases were in steps of 1/1.2 (or 0.83) times the previous length. Within each block of 100 trials, separate adaptive tracks were used for trials with valid and invalid primes, of which there were 50 for each trial type. Across all trials and all blocks, prime and target sentences were drawn without replacement from the experimental subset of the TIMIT database, and

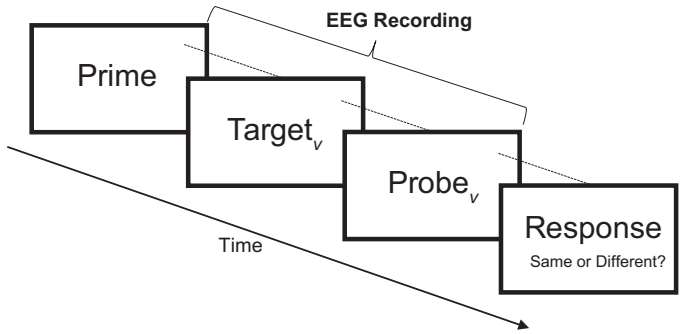

Fig. 1. Schematic of an experimental trial. A vocoded target (Target$_V$) is preceded by a natural speech (nonvocoded) cue (Prime) and followed by a vocoded probe (Probe$_V$). Participants are asked to indicate whether (same) or not (different) the probe snippet was drawn from the target (Response). On half the trials, the cue and the target were matched (valid), and on the other half they were mismatched (invalid).

sentences were drawn at random for each participant. EEG was recorded simultaneously with this task. Although we used an adaptive track in an attempt to avoid confounds of unequal effort across conditions, we did not employ an objective measure of attention.

EEG recording and preprocessing. High-density EEG (128 channels) was recorded with equipment from Neuroscan. Electrodes were placed following the international 10/5 system (Oostenveld and Praamstra 2001), and all channel impedances were kept below 10 kΩ. The EEG data were sampled at 1,000 Hz and filtered offline with a passband of 1–50 Hz. The filtered data were then segmented into individual trials which were 3 s long, beginning 500 ms after the start of the sentence. This delay was incorporated to remove the onset response to the start of the sentence. Artifacts were removed from the segmented EEG data using the Infomax ICA algorithm from the EEGLAB toolbox (Delorme and Makeig 2004).

Envelope extraction and cross-correlation analysis. The envelopes of the speech materials were extracted, bandpass filtered from 1 to 50 Hz, and downsampled to 1,000 Hz. The first 500 ms of this envelope were removed and the subsequent 3 s retained to align with the EEG data. This procedure was performed for both the natural speech primes and the vocoded targets. Paired with the neural response, this broadband speech envelope was submitted to a cross-correlation to quantify the entrainment response to speech. This cross-correlation was performed for both the vocoded target sentences and the natural speech prime sentences for all recording channels (Fig. 2). The cross-correlation function measures the similarity between two discrete signals f and g over a range of delays n, where m is discrete time:

$$(f \star g)(n) = \sum_{m=-\infty}^{\infty} \frac{f[m]g[n+m]}{std(f)std(g)}.$$

The cross-correlation functions between the EEG response and the stimulus envelope peak at an average latency of 75 ms, slightly earlier than the typical N1 response of an auditory evoked potential. For ease of discussion, however, we will refer to this peak as an N1, because it occurs within the N1 range and because the cross-correlation function is expected to reflect a contribution from typical N1 generators. For every trial, recordings from each channel of the EEG were cross-correlated with both the target and the prime sentence envelopes, and cross-correlation values were Fisher z-transformed to provide an approximately normal distribution, following the analysis in Baltzell et al. (2016).

Cross-correlation functions were then averaged across trials and subjects to form a grand average, and mean values were extracted in the N1 latency range for each channel, which extended from 25 to 125 ms. The absolute value of these means was computed so that a region of interest (ROI) could be defined without respect to polarity. We chose to analyze mean rather than peak cross-correlation values so as

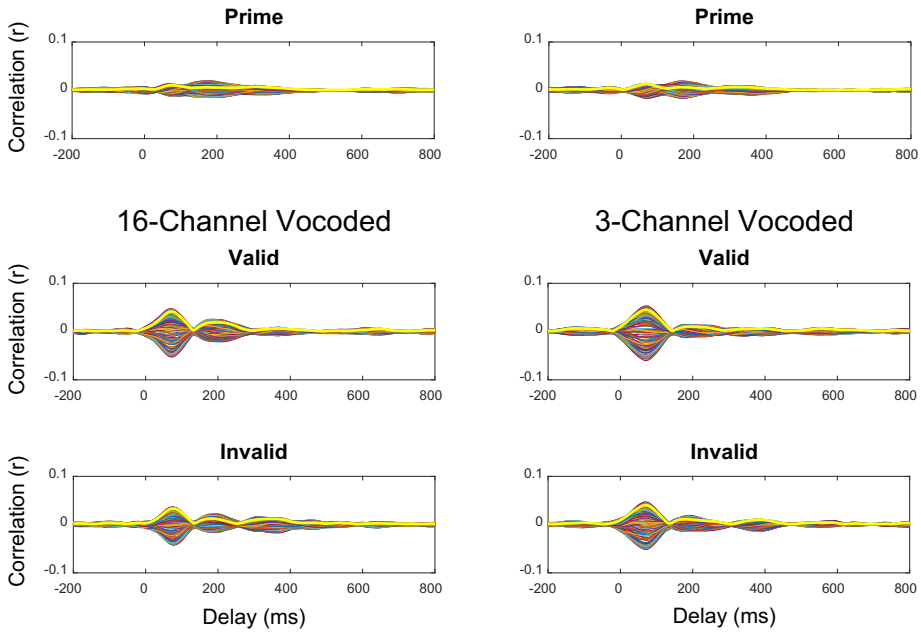

Fig. 2. Grand-averaged cross-correlation functions [bandpass-filtered envelopes (1–50 Hz) were cross-correlated with the EEG recorded from that trial] for the natural speech cues and the vocoded targets. Thick yellow lines indicate the time course of the ROI. The only prominent peak in these functions is contained within a latency range of ~25 to ~125 ms.

to correct for any noise in the estimation of the neural delay. This noise may be physiological, or it may be an artifact of the large variability of our speech stimuli. To define the ROI, we selected the 30 electrode channels with the highest mean cross-correlation in the N1 latency range for all sentence types. These six sentence types included the natural speech prime and the vocoded targets in the valid and invalid prime conditions (prime/valid/invalid) for both the 3-channel and 16-channel vocoder conditions. From this set of 180 (6 × 30) channels, duplicates were removed, leaving only 44 unique channels, and these 44 channels comprised our ROI (Fig. 3). Having defined our ROI from the grand-averaged data, for each subject, the mean of the absolute value of the cross-correlation functions for each sentence type was computed first over the ROI and then over the N1 latency range (25 to 125 ms).

To estimate the variability in these means, a bootstrap simulation was performed. A control distribution was constructed by replacing the sentences on each trial with randomly chosen sentences not presented on that trial and then performing the same analysis described above. This control was useful because it maintained the average spectral and temporal characteristics of the experimental sentences but was unrelated to the sentences used on a particular trial.

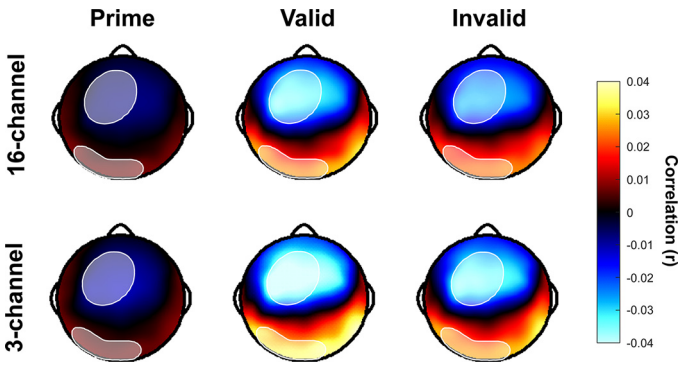

Fig. 3. Scalp topographies for the latency window of interest (25–125 ms). EEG was recorded with a 128-channel electrode cap and sampled at 1,000 Hz with an online average reference. The recordings were then bandpass filtered from 1 to 50 Hz. Thirty channels that showed the largest (regardless of polarity) cross-correlations were selected from the grand average for each of the 6 conditions and were combined to form an ROI of 44 channels, indicated by the shaded region.

Therefore, any nonzero values in the control cross-correlations would be due to chance. For each subject, this bootstrap was repeated 10 times, and grand-averaged means were computed 10,000 times from these bootstraps to create the control distribution. Mean values were considered significantly nonzero if they fell outside the 99th or 1st percentiles of this distribution.

RESULTS

Behavior. To ensure that our intelligibility manipulations had their intended effect, we gathered subjective intelligibility ratings before each recording block. Participants were asked to indicate on a scale from 1 to 6 how intelligible the target vocoded sentence was. Participants were asked not to consider the prime when making their intelligibility judgements. As shown in Fig. 4*A*, median intelligibility ratings across subjects for the 16-channel blocks were almost exclusively 6, the main exception being the invalid prime condition in the first block. Presumably, this reflects the fact that even highly intelligible vocoded speech can seem foreign when heard for the very first time, and we only gathered a single intelligibility rating for each prime condition (valid/invalid) before EEG blocks for the 16-channel condition. Median intelligibility ratings for the 3-channel blocks demonstrate a consistent difference between valid and invalid prime conditions across all four blocks. Median intelligibility ratings for the valid prime trials range between 4.5 and 5, whereas median intelligibility ratings for the invalid prime trials range between 1 and 2.

During experimental (recording) trials, participants were asked to indicate whether or not probe clips of vocoded material were drawn from the vocoded target (Fig. 1). The durations of these probes were adaptively varied according to a 2-down/1-up tracking procedure, and estimated durations required for 71% correct are shown in Fig. 4*B*. The resulting probed durations are referred to as probe thresholds. For the 16-channel vocoded blocks, probe thresholds are virtually the same for valid and invalid trials, with a mean duration across blocks of 174 ms for valid trials and 186 ms for invalid trials. This is consistent with intelligibility ratings that are also

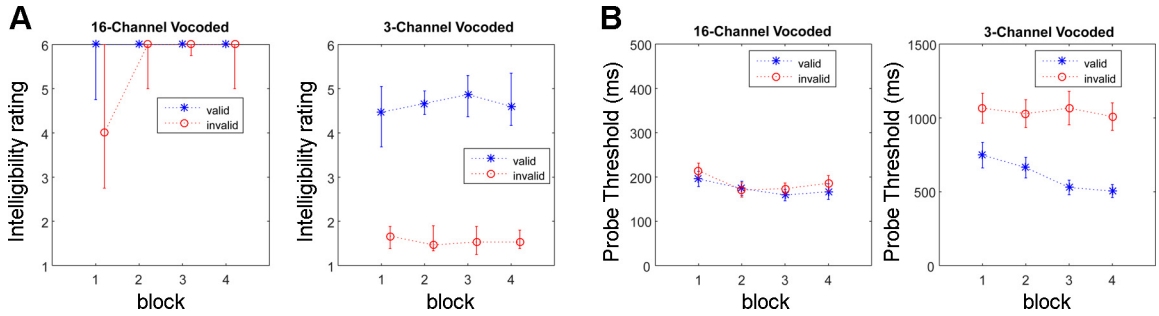

Fig. 4. Behavioral results. *A*: median subjective intelligibility ratings (25th and 75th quartiles shown) measured before each block. *B*: threshold probe durations for each block (note change in scale).

similar across valid and invalid trials. For the 3-channel vocoded blocks, probe thresholds are substantially higher for invalid trials than for valid trials, with a mean duration across blocks of 1,043 ms for invalid trials and 612 ms for valid trials. This is again consistent with intelligibility ratings that show substantial differences between valid and invalid trials.

As expected, these behavioral results suggest that intelligibility depends on the number of vocoded channels and indicate no evidence of practice effects on the intelligibility of vocoded speech over the course of the experiment. Specifically, 3-channel vocoded speech with invalid primes remained consistently unintelligible over the course of the experiment.

Electrophysiology. A latency range for the N1 peak was defined (25–125 ms) on the basis of grand-averaged data, and a subset of maximally responding channels was selected to form a region of interest (ROI) within this latency range (Fig. 2). Cross-correlation means were then selected from the ROI time series, and a bootstrap was performed to estimate a noise floor for cross-correlation means due to chance (Fig. 5). A two-factor repeated-measures ANOVA performed on the cross-correlation means failed to reveal a significant interaction between cue validity and number of vocoded channels [$F(1, 12) = 0.14$, $P = 0.714$]. Consistent with Millman et al. (2015) then, we failed to demonstrate a significant effect of intelligibility on the entrainment response. The ANOVA revealed a significant main effect of cue validity [valid vs. invalid: $F(1, 12) = 13.6$, $P = 0.03$], suggesting that prior knowledge leads to an increased entrainment response. The ANOVA failed to reveal a significant main effect of number of vocoded channels [16 vs. 3: $F(1, 12) = 2.7$, $P = 0.126$], suggesting that sensory detail, at least with regard to 16- and 3-channel vocoded speech, has no effect on the entrainment response. However, because the entrainment response to 3-channel vocoded speech is larger across all conditions than the entrainment response to 16-channel speech, the lack of significant main effect we report may reflect of a lack of statistical power.

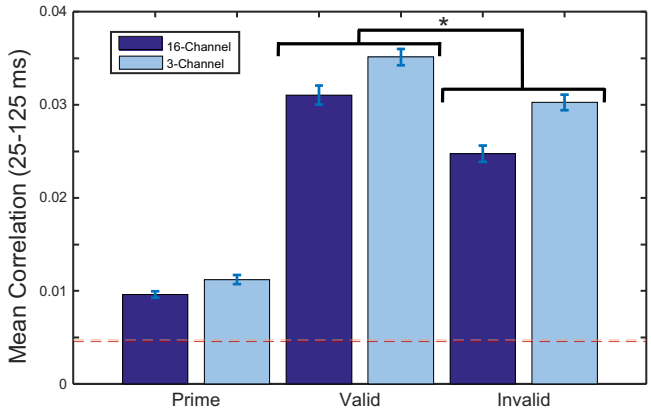

Fig. 5. Bar plots showing mean correlation values over the latency window (25–125 ms) and averaged across subjects. The bootstrapped noise floors (99th percentile cutoffs) are displayed as dotted lines (16-channel, red/dark; 3-channel, pink/light). A 2-factor repeated-measures ANOVA revealed a significant effect of valid vs. invalid [$F(1, 12) = 13.6$, $*P = 0.03$] but no significant effect of 16- vs. 3-channel vocoding [$F(1, 12) = 2.7$, $P = 0.126$] and no significant interaction [$F(1, 12) = 0.14$, $P = 0.714$]. Post hoc *t*-tests revealed a robust difference between the natural speech cue and the vocoded targets (all $P < 0.005$). This suggests that task demand may play a crucial role in determining the strength of the entrainment response.

When we compared the entrainment response to the vocoded targets and the entrainment response to the natural speech primes, we observed a significant difference. Post hoc *t*-tests revealed a significant difference between the natural speech cue and the vocoded targets (all $P < 0.005$, with a corrected critical value of 0.0125) such that the entrainment response to the natural-speech cue is significantly smaller than the entrainment response to vocoded targets. Effect sizes for all comparisons are greater than 1.7, suggesting that this effect is particularly robust. We chose to run a post hoc analysis on the clean speech primes rather than include them in the ANOVA for two reasons. First, the clean speech primes were not directly probed with a behavioral task and were therefore subjected to qualitatively different task demands than the vocoded speech. Second, our ANOVA was meant to address the effects of priming, sensory detail, and intelligibility, and as such we had decided that the analysis of the clean speech primes should be post hoc.

DISCUSSION

The data reported in this article suggest that *1*) the strength of the entrainment response to speech depends on prior knowledge, *2*) the strength of the entrainment response to speech depends critically on task demand (prime vs. target), and *3*) the strength of the entrainment response to speech does not depend on intelligibility, although a lack of significance should not necessarily be overemphasized, especially with a modest sample size ($n = 13$).

Prior acoustic knowledge modulates the entrainment response. We found that for both the 16-channel and 3-channel vocoded speech, the entrainment response in the valid prime condition is larger than the entrainment response in the invalid prime condition. This suggests that prior knowledge modulates the entrainment response. Furthermore, we did not find a significant difference between the entrainment response to 3-channel and 16-channel vocoded speech, which is to say we did not

observe an effect of sensory detail. These results are consistent with the results of Sohoglu et al. (2012), who found an effect of prior knowledge while failing to find an effect of sensory detail on the EEG response. However, whereas Sohoglu et al. (2012) found the most robust effects of prior knowledge in the 270- to 700-ms post-onset latency range, we did not observe a reliable entrainment response at these later latencies, and therefore our analysis was restricted to the 25- to 125-ms latency range. Nonetheless, the effects of prior knowledge and sensory detail on the entrainment response are broadly consistent with previously reported effects of prior knowledge and sensory detail on the auditory evoked response.

A lack of significant interaction between prior knowledge (valid vs. invalid) and sensory detail (16-chanel vs. 3-channel) suggests that the effect of prior knowledge is primarily acoustic. Because 16-channel speech is more intelligible than 3-channel speech, we would expect the effect of prior knowledge to be more substantial for 16-channel speech if we suppose that linguistic prior knowledge is modulating the entrainment response. However, if acoustic prior knowledge is modulating the entrainment response, we would expect no difference in the effect of prior knowledge across vocoding conditions, and this is what we observed.

Task demand modulates the entrainment response. Although we did not set out to explicitly test the effect of task demand on the entrainment response, our results suggest that the strength of the entrainment response depends on the task relevance of the stimulus. Specifically, we found that the entrainment response to the natural speech prime sentences was far smaller than the entrainment response to vocoded targets, regardless of condition. Considering that natural speech contains no artificial acoustic distortions, it seems unlikely that stimulus acoustics alone explain this observation. However, Ding et al. (2014) found that in the 1 - to 4-Hz range, the entrainment response to 4-channel vocoded speech was significantly larger than the entrainment response to natural speech, suggesting that even in the absence of unequal task demands, we might expect the entrainment response to natural speech to be smaller than the entrainment response to vocoded speech. Tone-vocoding distorts the fine structure of natural speech, which may cause listeners to deploy more resources to the encoding of the speech envelope and less to the encoding of the fine structure. For natural speech, the inverse may be true and could therefore help explain why the entrainment response is smaller to the natural speech cue than to the vocoded target. In Fig. 5, however, we show that the entrainment response to the natural speech primes is substantially reduced relative to the vocoded targets, beyond what is expected from the results of Ding et al. (2014). This suggests that the difference in stimuli alone is not sufficient to explain our result. Instead, it suggests that task demand may play a crucial role in determining the strength of the entrainment response, reinforcing the importance of adequately controlling behavioral tasks across intelligibility conditions.

In other words, the fact that listening closely to the natural speech primes was not necessary to perform the behavioral task may be responsible for the difference in entrainment to the natural speech primes and the vocoded targets. Participants were asked to compare probe vocoded clips to the target vocoded sentence, which requires acute attention to the vocoded target. On the other hand, whereas listening to the natural speech prime affects the intelligibility of the vocoded target on valid trials, acute attention is not required. Although it is well known that attention modulates the entrainment response in multitalker listening scenarios (e.g., Ding and Simon 2012; Horton et al. 2013), our results suggest that attention is crucial even in the absence of a competing talker. However, given that we did not explicitly control for task demand across listening to the natural speech primes and vocoded targets, further studies that explicitly control for and quantify attentional demands are needed.

Intelligibility does not modulate the entrainment response. Following Millman et al. (2015), we used priming to study the effect of intelligibility on the entrainment response to vocoded speech, and consistent with Millman et al. (2015), we failed to find an effect of intelligibility. Crucially, we failed to find an effect of intelligibility using novel speech stimuli on each trial and an attentionally demanding behavioral task that was adaptively varied to achieve equal percent correct across conditions. Thus our result is consistent with previous results suggesting that entrainment to naturalistic speech envelopes is driven by acoustic rather than linguistic neural processes (Howard and Poeppel 2010; Millman et al. 2015; Zoefel and VanRullen 2016).

A potential confound in our study, however, is the fact that our behavioral task explicitly directed listeners to the acoustics of the speech stimulus, and it is possible that if we had used a task that more explicitly focused on the semantic and/or syntactic aspects of the stimulus, we might have observed an effect of intelligibility. However, probe duration thresholds are shortest for 16-channel vocoded speech, are longer for validly primed 3-channel vocoded speech, and are still longer for invalidly primed 3-channel speech, thus neatly scaling with subjective intelligibility reports (Fig. 4). This suggests that listeners are making use of linguistic rather than strictly acoustic information when comparing the probe to the target, at least to the extent that "linguistic" refers to those overlearned features of the clean speech stimulus that allow for the restoration of intelligibility to degraded speech.

Another potential confound is the fact that the effect of prior knowledge might not be equal across vocoding conditions. Because 16-channel vocoded speech more closely resembles natural speech than 3-channel vocoded speech, we might expect the effectiveness of the prime to be greater for 16-channel speech. If the prime was more effective at modulating the entrainment response in the 16-channel vocoded condition, it could mask a subtle effect of intelligibility in the 3-channel condition. If validly primed 3-channel vocoded speech is modulated by both prior knowledge and intelligibility, and validly primed 16-channel vocoded speech is modulated only by prior knowledge, we would expect to see a significant interaction between prior knowledge and sensory detail. However, if prior knowledge is disproportionately effective at modulating the entrainment response to 16-channel vocoded speech, we might fail to observe this interaction. Nonetheless, our data are consistent with literature suggesting that the entrainment response is primarily driven by acoustic properties of the speech signal.

One hypothesis, put forward by Doelling et al. (2014), is that entrainment in the theta band is necessary but not sufficient for speech perception. This hypothesis is based on the finding that the magnitude of abrupt changes at the start of syllables are

predictive of syllable length (Greenberg et al. 2003). It is possible, then, that neural entrainment is driven by this initial syllabic chunking mechanism, which is acoustic in origin but may help support subsequent linguistic processes.

Interactions among multiple sources. It is almost certainly the case that the scalp-recorded entrainment response we report is the resulting sum of multiple underlying sources that are driven by different aspects of the speech stimulus. However, because we could not resolve distinct sources that were consistent across subjects in this data set, we report only the sum of these sources. It is therefore possible that sources driven by acoustic properties of the speech are masking a source that is modulated by intelligibility.

Conclusions. Whereas the strength of the entrainment response to speech depends on prior knowledge, we failed to find evidence that the entrainment response depends on intelligibility. However, a number of potential confounds make it difficult to rule out the possibility that our design masked a subtle effect of intelligibility. Finally, the strength of the entrainment response to speech appears to critically depend on task demand, underscoring the importance of controlling task demand in entrainment studies.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

L.S.B., R.S., and V.M.R. conceived and designed research; L.S.B. performed experiments; L.S.B. analyzed data; L.S.B., R.S., and V.M.R. interpreted results of experiments; L.S.B. prepared figures; L.S.B. drafted manuscript; L.S.B., R.S., and V.M.R. edited and revised manuscript; L.S.B., R.S., and V.M.R. approved final version of manuscript.

REFERENCES

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98: 13367–13372, 2001. doi:10.1073/pnas.201400998.

Baltzell LS, Horton C, Shen Y, Richards VM, D'Zmura M, Srinivasan R. Attention selectively modulates cortical entrainment in different regions of the speech spectrum. *Brain Res* 1644: 203–212, 2016. doi:10.1016/j.brainres.2016.05.029.

Buchweitz A, Mason RA, Tomitch LM, Just MA. Brain activation for reading and listening comprehension: An fMRI study of modality effects and individual differences in language comprehension. *Psychol Neurosci* 2: 111–123, 2009. doi:10.3922/j.psns.2009.2.003.

Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. *Science* 304: 1926–1929, 2004. doi:10.1126/science.1099745.

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight RT. High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313: 1626–1628, 2006. doi:10.1126/science.1128115.

Cohen L, Lehéricy S, Chochon F, Lemer C, Rivaud S, Dehaene S. Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain* 125: 1054–1069, 2002. doi:10.1093/brain/awf094.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134: 9–21, 2004. doi:10.1016/j.jneumeth.2003.10.009.

Ding N, Chatterjee M, Simon JZ. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88: 41–46, 2014. doi:10.1016/j.neuroimage.2013.10.054.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19: 158–164, 2016. doi:10.1038/nn.4186.

Ding N, Simon JZ. Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8: 311, 2014. doi:10.3389/fnhum.2014.00311.

Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107: 78–89, 2012. doi:10.1152/jn.00297.2011.

Doelling KB, Arnal LH, Ghitza O, Poeppel D. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85: 761–768, 2014. doi:10.1016/j.neuroimage.2013.06.035.

Fries P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn Sci* 9: 474–480, 2005. doi:10.1016/j.tics.2005.08.011.

Fries P, Reynolds JH, Rorie AE, Desimone R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291: 1560–1563, 2001. doi:10.1126/science.291.5508.1560.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N, Zue V. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium, 1993.

Ghitza O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol* 2: 130, 2011. doi:10.3389/fpsyg.2011.00130.

Giraud AL, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15: 511–517, 2012. doi:10.1038/nn.3063.

Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech—a syllable-centric perspective. *J Phon* 31: 465–485, 2003. doi:10.1016/j.wocn.2003.09.005.

Hertrich I, Dietrich S, Ackermann H. Tracking the speech signal–time-locked MEG signals during perception of ultra-fast and moderately fast speech in blind and in sighted listeners. *Brain Lang* 124: 9–21, 2013. doi:10.1016/j.bandl.2012.10.006.

Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. *J Neurophysiol* 109: 3082–3093, 2013. doi:10.1152/jn.01026.2012.

Howard MF, Poeppel D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol* 104: 2500–2511, 2010. doi:10.1152/jn.00251.2010.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320: 110–113, 2008. doi:10.1126/science.1154735.

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94: 1904–1911, 2005. doi:10.1152/jn.00263.2005.

Levitt H. Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49, *Suppl* 2: 467–477, 1971. doi:10.1121/1.1912375.

Loizou PC, Dorman M, Tu Z. On the number of channels needed to understand speech. *J Acoust Soc Am* 106: 2097–2103, 1999. doi:10.1121/1.427954.

Millman RE, Johnson SR, Prendergast G. The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J Cogn Neurosci* 27: 533–545, 2015. doi:10.1162/jocn_a_00719.

Nie K, Stickney G, Zeng F-G. Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Trans Biomed Eng* 52: 64–73, 2005. doi:10.1109/TBME.2004.839799.

Nunez PL, Srinivasan R. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford: Oxford University Press, 2006.

Oostenveld R, Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clin Neurophysiol* 112: 713–719, 2001. doi:10.1016/S1388-2457(00)00527-7.

Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23: 1378–1387, 2013. doi:10.1093/cercor/bhs118.

Price CJ. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62: 816–847, 2012. doi:10.1016/j.neuroimage.2012.04.062.

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. *Science* 212: 947–949, 1981. doi:10.1126/science.7233191.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science* 270: 303–304, 1995.

Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive top-down integration of prior knowledge during speech perception. *J Neurosci* 32: 8443–8453, 2012. doi:10.1523/JNEUROSCI.5069-11.2012.

Thompson LW, Thompson VD. Comparison of EEG changes in learning and overlearning of nonsense syllables. *Psychol Rep* 16: 339–344, 1965. doi:10.2466/pr0.1965.16.2.339.

Zoefel B, VanRullen R. EEG oscillations entrain their phase to high-level features of speech sound. *Neuroimage* 124, Pt A: 16–23, 2016. doi:10.1016/j.neuroimage.2015.08.054.