

Research Article

Speech Perception with Noise Vocoding and Background Noise: An EEG and Behavioral Study

YUE DONG¹, AND YAN GAI¹ 

¹ Biomedical Engineering, Parks College of Engineering, Aviation and Technology, Saint Louis University, 3507 Lindell Blvd, St Louis, MO 63103, USA

Received: 13 March 2020; accepted: 26 January 2021; Online publication: 13 April 2021

ABSTRACT

This study explored the physiological response of the human brain to degraded speech syllables. The degradation was introduced using noise vocoding and/or background noise. The goal was to identify physiological features of auditory-evoked potentials (AEPs) that may explain speech intelligibility. Ten human subjects with normal hearing participated in syllable-detection tasks, while their AEPs were recorded with 32-channel electroencephalography. Subjects were presented with six syllables in the form of consonant-vowel-consonant or vowel-consonant-vowel. Noise vocoding with 22 or 4 frequency channels was applied to the syllables. When examining the peak heights in the AEPs (P1, N1, and P2), vocoding alone showed no consistent effect. P1 was not consistently reduced by background noise, N1 was sometimes reduced by noise, and P2 was almost always highly reduced. Two other physiological metrics were examined: (1) classification accuracy of the syllables based on AEPs, which indicated whether AEPs were distinguishable for different syllables, and (2) cross-condition correlation of AEPs (r_{cc}) between the clean and degraded speech, which indicated the brain's ability to extract speech-related features and suppress response to noise. Both metrics decreased with degraded speech quality. We further tested if the two metrics can explain cross-subject variations in their behavioral performance. A significant correlation existed for r_{cc} , as well as classification based on early AEPs, in the fronto-central areas. Because r_{cc} indicates similarities between clean and degraded speech, our

finding suggests that high speech intelligibility may be a result of the brain's ability to ignore noise in the sound carrier and/or background.

Keywords: EEG, noise vocoding, auditory evoked potentials, noise suppression, speech intelligibility, speech perception

INTRODUCTION

Human speech contains highly redundant information (Stilp 2011; Beekhuizen et al. 2013). When noise is introduced to sound carriers or added in the background, speech perception can stay high until too much distortion occurs (Eisenberg et al. 2000; Friesen et al. 2001; Shannon et al. 2004). One may ask the following questions: Does the brain activity reflect the distortion when speech is still highly intelligible? What features of the brain signals may be used to predict human's speech perception?

In this study, we examined two features of brain signals recorded from human subjects who were performing syllable-recognition tasks. Ten normal-hearing subjects were presented with clean or noise-vocoded syllables in the form of consonant-vowel-consonant or vowel-consonant-vowel, while their auditory-evoked potentials (AEPs) were recorded by a 32-channel electroencephalography (EEG) system. In some conditions, background noise was added to the non-vocoded or vocoded speech. Subjects were asked to select the perceived syllable from a list of six.

First, we examined traditional AEP features as positive and negative peaks, such as the P1, N1, and P2. The P1-N1-P2 complex was frequently used in previous studies using clean (Digeser et al. 2009; Swink and Stuart

Correspondence to: Yan Gai · Biomedical Engineering, Parks College of Engineering, Aviation and Technology · Saint Louis University · 3507 Lindell Blvd, St Louis, MO, 63103, USA. email: yan.gai@slu.edu

2012; Wagner et al. 2016), vocoded (Friesen et al. 2009; Won et al. 2011; Utianski et al. 2015), and noisy (Koerner et al. 2016; Faucette and Stuart 2017) speech. In general, the amplitudes of N1 and P2 are typically reduced with the addition of background noise (Faucette and Stuart 2017). It has also been shown that the addition of background noise significantly reduced a “P3” component (Koerner et al. 2016), which agrees with our P2 in terms of the latency. However, it was suggested that the behavior of the P1-N1-P2 complex with vocoded speech cannot fully account for human perception (Friesen et al. 2009).

Therefore, apart from those peaks, two other EEG metrics were also tested. The first was the classification accuracy for individual syllables. That is, after a classifier has been trained, can it predict which of the six syllables was presented given a particular brain waveform? The rationale of using this metric is that, when the subject can correctly choose the perceived syllable, his/her brain signals must respond differentially to individual syllables in a consistent way. In contrast, when speech is highly distorted, especially when random noise is introduced to the sound carrier or background, brain signals may no longer be able to consistently differentiate individual syllables. Therefore, one would predict that the classification performance should decrease as the listening condition becomes more difficult. Similar classification studies have been done previously with non-vocoded speech (Brandmeyer et al. 2013; Kim et al. 2014; Mugler et al. 2014).

However, one could also argue that a better classification may not indicate better speech recognition, if the brain waveforms to individual syllables are highly distorted. When our brain can correctly perceive distorted speech, it should be able to extract speech-related features and more or less suppress its response to noise in the vocoder or background, thus making it more similar to its clean-speech response. Previous studies showed that speech envelopes are found to be linearly related to EEG within 2–8 Hz (Pasley et al. 2012; O’Sullivan et al. 2015). The mechanism probably originates from *cortical entrainment*, which refers to phase-locked brain activity to speech envelope and other linguistic features. Cortical entrainment is believed to play an important role in speech perception (Peelle et al. 2013; Ding et al. 2014; Weisz and Obleser 2014; Baltzell et al. 2017; Puvvada et al. 2017; Steinmetzger and Rosen 2017). In particular, some studies focused on examining the so-called “noise-robust” cortical entrainment to speech envelope (Ding et al. 2014; Puvvada et al. 2017).

Although the present study did not directly examine cortical entrainment, the concept predicts that, if the brain does suppress its response to noise when being entrained to speech envelope, AEPs evoked by vocoded speech and/or speech in background noise should more or less resemble the clean-speech AEPs. Therefore, the second metric we examined was the cross-condition correlations of AEPs between clean and vocoded speech

with/without background noise. A high correlation value would indicate that the brain is capable of ignoring noise in the vocoder or background.

For the two physiological metrics, we further examined if one of them can explain cross-subject behavioral variations. In other words, when a human subject achieves high speech intelligibility with severely distorted speech, does the person also show highly distinguishable AEPs for different syllables, or, alternatively, high similarities of AEPs between the clean and distorted speech?

Methods

Human Subjects and EEG Recording

Ten human listeners (aged 19–40 years; four females and six males) participated in the study. The listeners were screened to have pure-tone thresholds of ≤ 20 dB HL at 0.25, 0.5, 1, 2, and 4 kHz. Experiments were conducted in a double-walled sound booth (8' \times 8' \times 8.1'; Noise Barriers LLC). The subjects performed behavioral tasks while their EEG signals were recorded. The experimental protocol was approved by the Institutional Review Board of Saint Louis University.

EEG signals were obtained with a 32-channel portable system (eegoTM sports; ANT Neuro) including a head cap, an amplifier, and a Windows tablet computer. There were 31 active electrodes using the common average (i.e., the average of all the electrodes was subtracted from each electrode recording). Each electrode cup was filled with conductive gel (Onestep Cleargel). The sampling frequency of the EEG amplifier was 500 Hz. Diotic sound was generated with a sampling frequency of 44.1 kHz using MATLAB (MathWorks) and delivered to a StimTracker (Cedrus). The StimTracker relayed the sound to a headphone (Sennheiser HD 280 PRO), while sending precisely timed markers to the EEG amplifier. Sound levels were calibrated with a Sound Pressure Meter (Extech).

An independent component analysis approach (Zhou and Gotman 2009) was applied offline to remove blinking artifacts. About 2% of trials (51 out of 2700) were then rejected from the analysis, since those trials contained notable artifacts. To obtain AEPs, signals were bandpass-filtered between 0.1 and 20 Hz using a 50th-order finite impulse response filter. The signals were then averaged over 75 trials for each syllable and sound condition to obtain an AEP.

Sound and Behavioral Paradigms

A total of six syllables (i.e., /aba/, /aka/, /asa/, /hæd/, /hɒd/, and /hid/) were tested using combined consonant-vowel syllables. Consonant-centered syllables spoken by a male talker were presented in the form of /aCa/, with

C being one of the three: $/b, k, s/$ (Shannon et al. 1999). Vowel-centered syllables spoken by a female talker were presented in the form of $/hVd/$, with V being one of the three: $/æ, \alpha, i/$ as in “had”, “hod”, and “heed” (Hillenbrand et al. 1995). Figure 1b shows example spectrograms of non-vocoded (top) and 22-channel vocoded syllables (bottom). The syllables started at time 0.

A noise-vocoding technique similar to Shannon et al. (1995) was applied to speech in quiet and in noise. Briefly, the number of bandpass filters was 22 or 4, with center frequencies evenly spaced on a logarithmic scale from 0.5 to 10 kHz. The filters were implemented as 3rd-order Butterworth filters, and the bandwidth of each filter was $24.7 + 0.1CF$ (Hz), according to the equivalent rectangular bandwidth (ERB) of human listeners with normal hearing (Glasberg and Moore 1990). The low-pass filtering for envelope extraction had a cutoff frequency of 150 Hz. The noise carrier was a broadband noise passing through the same bandpass filters and was randomly generated for each filter. As can be seen in the spectrograms, noise vocoding smeared the spectral features (Fig. 1b, bottom). The final sound level of the natural or vocoded speech alone without background noise was approximately 50 dB SPL, i.e., variations of the number of channels did not affect the overall sound level.

One difference between our approach and the Shannon et al. study is that, when the channel number was decreased to four, we did not broaden the bandwidths but kept them as the human ERB. The rationale is that, when the bandwidth broadens, the envelope after

bandpass filtering will be altered. For example, after the vocoding process, we compared the narrowband (e.g., 50 Hz) envelopes between clean and vocoded speech and found a correlation value of 0.7 using our ERB bandwidths. The correlation using the continuous bandwidths was only 0.4. As a consequence, our approach did not cover the entire frequency range.

The background noise was a white noise of 33 dB SPL with a flat spectrum between 0.4 and 15 kHz. The noise waveform was randomly created on each trial and was independent from the noise carrier during vocoding. Onset of the background noise was 0.2 s prior to the speech onset. When background noise was added, the speech level was lowered from 50 to 40 dB SPL so that a significant amount of behaviorally incorrect responses may be obtained with the 4-channel condition, yielding a signal-to-noise ratio (SNR) of 7 dB. The speech and broadband noise, if any, were added together before noise vocoding was applied.

Note that we chose a moderate SNR because large effects of background noise can already be seen at this level. Although, behaviorally, a lower SNR may generate more detection errors with the human subjects, it would be difficult to extract speech-related information using EEG (a better recording method, such as invasive electrode arrays, may be more suitable to perform the task).

There were three vocoding conditions: non-vocoded and vocoded with 22 or 4 channels. For each vocoding condition, there was either background noise or not. Therefore, a total of six conditions were obtained from each subject. Because there were six syllables and six sound conditions, a total of $6 \times 6 \times 75 = 2700$ trials were obtained from each subject, with 2% of the trials further rejected for containing artifacts.

The stimulus condition (vocoded vs. non-vocoded, in quiet vs. in noise) was fixed for a given recording session with the six syllables presented in a random order. On each trial, a syllable with or without background noise was randomly selected from the six syllables, and the subjects were asked to choose the perceived syllable from the list. Trials were presented every 7 s, and a behavioral/recording session contained 90 trials. Because sound presentations were clearly different for vowel- and consonant-centered syllables, the subjects were essentially choosing from three options in each group, and thus, the chance performance was $1/3$. A correct percentage (\hat{p}) was obtained in the end.

To test the significance of behavioral performance, the confidence interval, CI , was computed as

$$CI = \pm 1.960 \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (1)$$

(Lock et al. 2017) for a 95% confidence interval. If $CI - \hat{p}$ was greater than chance (i.e., $1/3$), a statistical significance was reached. When comparing two percentages,

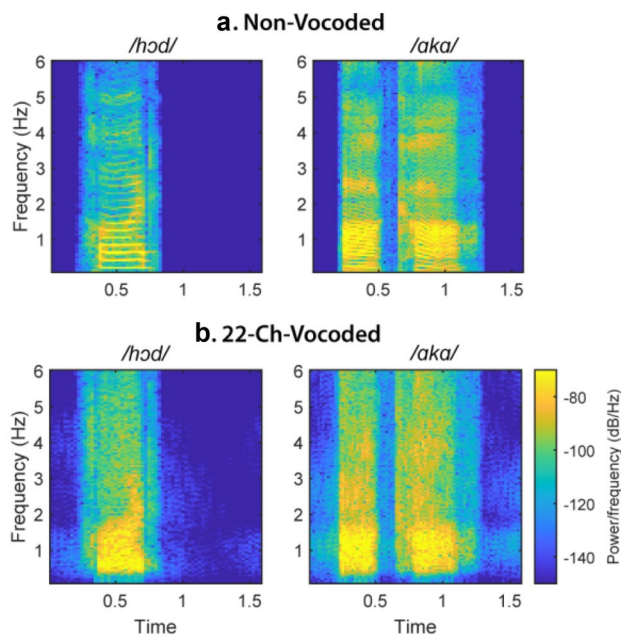


Fig. 1 **a** Spectrograms of two non-vocoded syllables, $/hod/$ and $/aka/$. **b** Spectrograms of the two syllables after being vocoded with 22 frequency channels

p_1 and p_2 , a two-proportion z test (Lock et al. 2017) can be computed as

$$z = (p_1 - p_2) / \sqrt{p_1(1 - p_1)/n + p_2(1 - p_2)/n} \quad (2)$$

with the null hypothesis being $H_0 : p_1 = p_2$. Here, $n = 75 \times 6 = 450$ trials for all six syllables in each sound condition.

AEP Peak Analysis

We examined three peaks that were most prominent in the AEPs, namely, the P1, N1, and P2. Not every electrode showed all three peaks, and not every subject showed the same peaks. Therefore, in the result (Table 1), we only presented the peaks that occurred consistently for the majority of subjects for each electrode.

Because different syllables can create variations in the peak time, a relatively large window, 200 ms, was used in finding the peak value. Specifically, after the mean peak latencies (Table 1) had been identified, a window of ± 200 ms was used to find the maximum or the minimum value of the peak. Note that this large window may result in overlaps in the adjacent windows. Fortunately, for P1, N1, and P2, we were computing the maximum, minimum, and maximum of the EEG traces in those

windows, respectively. Therefore, the overlaps should not cause confusions in the result.

Note that P2 would remain P2, even if there was no discernible P1 before. Also, as mentioned later, there can be a very early negative peak around 0.1 or 0.2 s caused by the onset of background noise. This peak was excluded from the peak analysis.

EEG Signal Classification

Signal classification measures how responses to different syllables were *consistently different* from one another. A linear discriminant analysis (LDA) (Balakrishnama et al. 1999) was used to classify AEPs elicited by different syllables. We examined the early (0–0.5 s) and late portions (0.5–1 s) of the AEPs separately for classification accuracy.

A “leave-one-out-cross-validation” approach based on combined trials was used to provide the training and test samples to the classifier. Each time, four trials obtained with the same class (i.e., syllable) were randomly selected and averaged to form a test waveform. Meanwhile, 200 training waveforms for each class were randomly formed from the pool, excluding trials that had been selected as the test waveform. A 1D weighting function was derived using the LDA and applied to the test waveform to come

TABLE 1

A, mean peak latencies for all the electrodes when there were consistent peaks across the human subjects. NA, a consistent peak cannot be identified for the majority of subjects

Electrodes	Peak Latencies (s)			Electrodes	Peak Latencies (s)		
	P1	N1	P2		P1	N1	P2
Fp1	NA	NA	NA	C4	NA	0.20	0.28
Fpz	NA	NA	NA	T8	NA	NA	NA
Fp2	NA	NA	NA	M2	0.20	0.28	NA
F7	NA	NA	NA	CP5	NA	0.27	NA
F3	NA	NA	NA	CP1	NA	0.20	0.28
Fz	NA	0.20	0.28	CP2	NA	0.20	0.30
F4	NA	0.20	0.28	CP6	NA	NA	NA
F8	NA	NA	NA	P7	0.20	0.28	NA
FC5	NA	NA	NA	P3	NA	NA	NA
FC1	0.15	0.20	0.28	Pz	NA	NA	NA
FC2	NA	0.20	0.27	P4	NA	NA	NA
FC6	NA	NA	NA	P8	0.18	0.28	NA
M1	0.20	0.30	NA	POz	NA	NA	NA
T7	0.20	0.28	NA	O1	0.18	0.28	NA
C3	NA	0.20	0.28	O2	0.20	0.28	NA
Cz	NA	0.20	0.30				

up with a decision variable. This procedure was repeated for 50 times to obtain the accuracy, \hat{p} . Because the structures of the syllables were different for vowels and consonants (Fig. 1a), a two-step classification was performed. The early or late AEP on each trial was first classified as a vowel or a consonant using the LDA. A second classification of the exact syllable was then performed within the vowel or consonant group using the LDA again, regardless of whether the group had been correctly determined in the previous step. We found that the two-step classification generated higher classification accuracies than a direct classification of six syllable types.

The \hat{p} value indicates, for a certain sound type, how consistently distinguishable the AEPs were with different syllables. By comparing \hat{p} values obtained with different sound conditions, the effects of noise vocoding and background noise can be examined. The early- and late-AEP \hat{p} value was further compared with human behavior in terms of cross-subject variations to answer the following question: *Is high speech intelligibility achieved together with distinguishable AEPs in the brain signals?*

EEG Cross Condition Correlations

Because there were six syllables, each sound condition elicited six AEPs, each averaged over 75 trials. For a given electrode, AEPs of the six syllables were concatenated in time to enable the computation of a single correlation value. This is a better approach than averaging individual-syllable correlations because differences in the AEP amplitude for different syllables can thus contribute to the concatenated correlation value. When obtaining the cross-condition correlation, r_{cc} , correlations were always made between the AEP in the control condition (non-vocoded syllables without background noise) and the AEP in a vocoded/noisy condition. The first 0.3 s of response was excluded from the correlation, since it may contain noise onset-induced activity irrelevant to the speech.

The resulting correlation indicates the *similarity* between the control AEP and a distorted sound AEP. This correlation value was further compared with human behavior to answer the following question: *Is high speech intelligibility achieved together with a high waveform similarity to the clean-speech response in the brain signals?*

Results

Behavioral Results

Figure 2 shows behavioral accuracies for individual subjects (A) and individual syllables (B). Results are grouped based on the sound condition, with black bars showing the background noise-free performance and gray bars showing the speech-in-noise performance. Error bars are

95% confidence intervals. For all subjects and syllables, the non-vocoded and 22-channel-vocoded conditions always showed near-perfect performances, whether or not accompanied with background noise.

With four vocoding channels, no-noise performance remained high for subjects S1, S2, S3, and S9 (Fig. 2a). For the rest of the subjects, lower performance was caused only by incorrect detections of /hæd/ and /hɔd/, as shown in individual performance (Fig. 2b, 3rd black bars). When background noise was added to four-channel vocoded speech, all subjects showed significantly decreased performance (Fig. 2a, asterisks; z test, with z , p , and df values specified in the figure). When examining the individual syllables, /hid/ had the highest performance; /hæd/ and /hɔd/ were still the lowest (Fig. 2b, 3rd gray bars). The result generally agreed with previous studies in that high speech recognition scores can be obtained with as few as four vocoding channels in easy listening conditions, but more channels are required for listening in noise (Shannon et al. 1995; Loizou et al. 1999; Friesen et al. 2001; Smith et al. 2002).

AEPs and Speech-Related Features

Figure 3a shows examples of AEPs for non-vocoded speech in quiet, each averaged over 75 trials, obtained with electrode FC1. Time 0 was the beginning of syllable presentation. Auditory potentials evoked by non-vocoded speech were highly fluctuating with multiple positive and negative peaks. The most consistent features were P1, N1, and P2 (Fig. 3a, left). When examining individual electrodes, all these three peaks may not always occur. Table 1 lists the peak latencies for all the electrodes. Here, we only identified those peaks that occurred consistently for the majority of subjects.

In the examples shown in Fig. 3, when random background noise was introduced to clean speech with an onset at -0.2 s (Fig. 3b, the beginning of the plot), noise-induced onset activity occurred around 0 to 0.1 s (Fig. 3b, arrow), which was not present in the noise-free responses (Fig. 3a). Meanwhile, the P1 and N1 associated with the speech signal (Fig. 3a) were clearly altered (Fig. 3b). With 22 vocoding channels but no background noise (Fig. 3c), most subjects showed similar AEPs to the clean-speech AEPs.

To quantify the effect of vocoding and background noise on AEPs, the absolute values of the P1, N1, and P2 were extracted. Figure 4 shows examples for electrode FC1. The x -axis is always the peak height without vocoding or background noise, serving as a control to clean speech. In A, the y -axis is the peak height for vocoded speech. Each symbol represents the average value for a single syllable over 75 trials, and hence, there are six symbols per condition for each subject with a total of 10 subjects. Overall, vocoding did not consistently change any

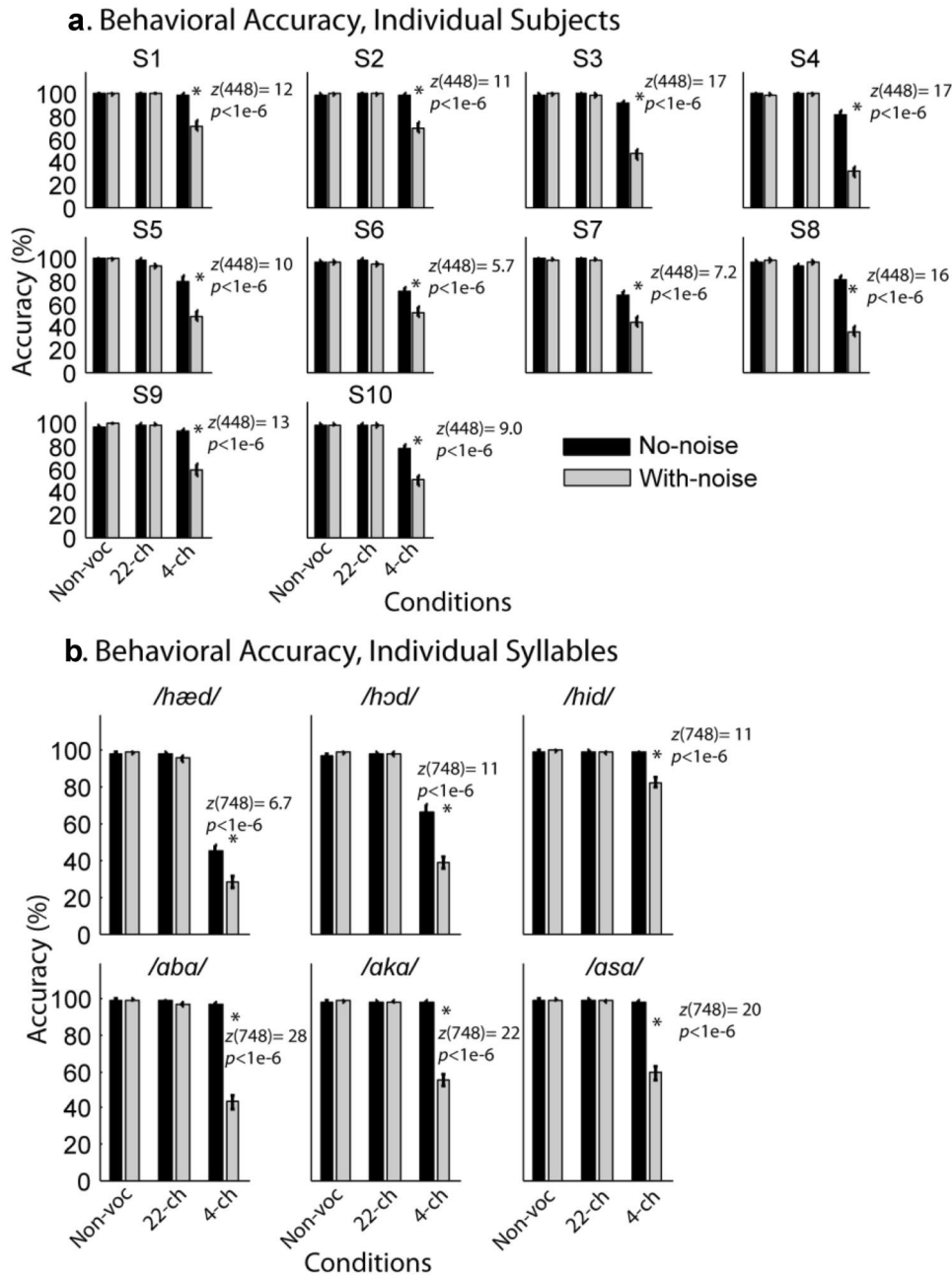


FIG. 2 Behavioral accuracies under different vocoding and noisy conditions averaged across all the syllables for each subject **a** or averaged across the subjects for each syllable **b**. Error bars are 95% confidence intervals. * marks a significant decrease when background noise was added to the 4-channel vocoded syllables (z test)

of the peaks. In contrast, the effect of background noise was more consistent in suppressing the speech-related P2 (Fig. 4b, rightmost). Most subjects and responses showed reduced peaks with background noise for electrode FC1 compared with the clean-speech condition. Note that the boosting effect on P1 (Fig. 4b, leftmost) was not real; it was due to the noise-onset activity (Fig. 3b).

When we examined P1, N1, and P3 for all the electrodes, vocoding alone rarely showed a consistent effect on P1, N1, or P2 (not shown). Therefore, in Fig. 5, only

the noisy conditions were plotted. Figure 5a shows effect of noise on P1; no electrode showed a significant reduction (*t* test; the *t*, *df*, and *p* values are specified in the figure); again, the boosting effect due to noise onset was omitted. For N1, more electrodes showed significant reductions (Fig. 5b). The largest reduction was observed for electrodes having a P2 (Fig. 5c). In fact, all the electrodes showing a prominent P2 had a significant reduction in their peak height with the addition of background noise.

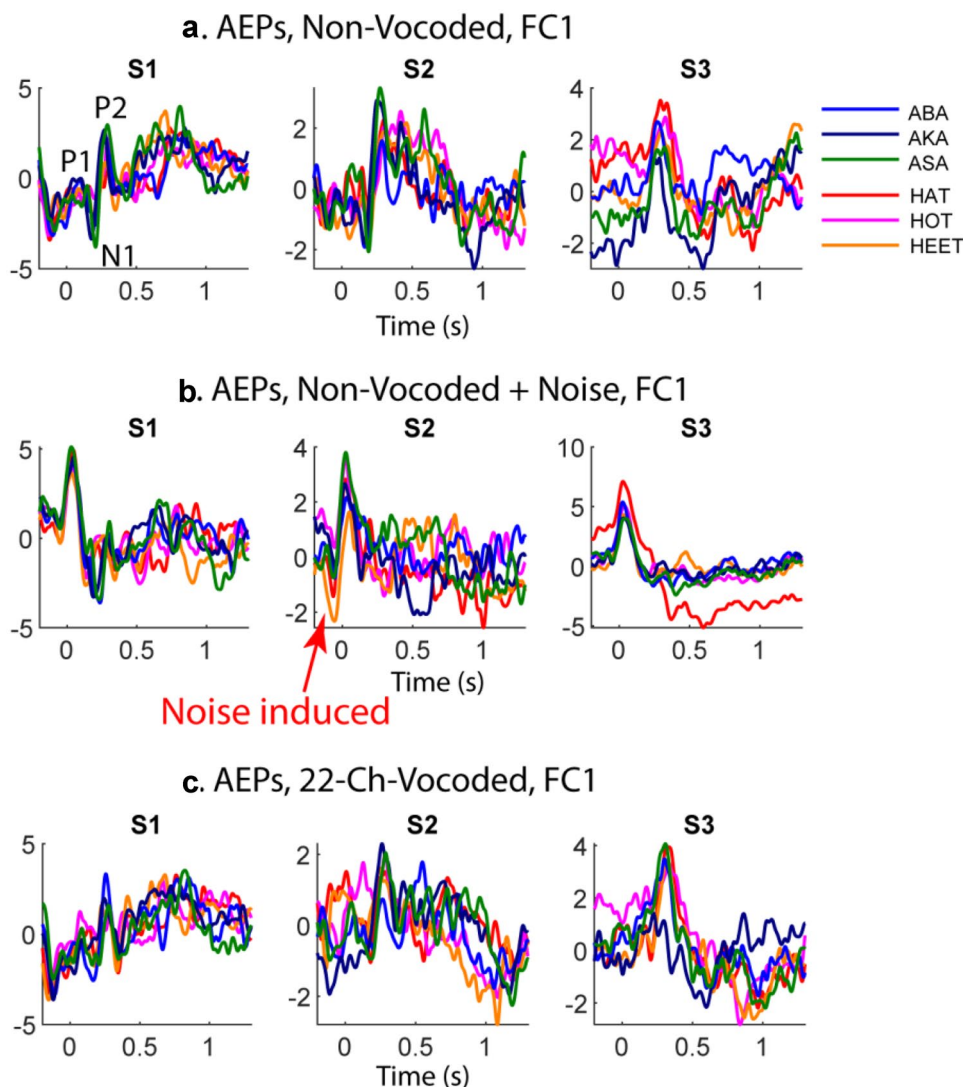


FIG. 3 Examples of AEPs obtained with electrode FC1 using non-vocoded (**a**, **b**) and vocoded (**c**) syllables for three subjects. In **b**, background noise was added to non-vocoded syllables. Note that speech began at time 0, whereas onset of the background noise

was 0.2 s prior to the speech onset, which was the beginning of the plot. The arrow pointed at noise-induced onset activity. Each AEP was an average over 75 trials

Vocoding and Noise on Classification Accuracy

Figure 6a shows examples of the scalp topography of signal classification accuracies based on the early AEPs (0–0.5 s) for four subjects and different sound conditions. Recall that a high accuracy indicates that AEPs elicited by the six different speech syllables were consistently different and, therefore, can be distinguished. Chance performance for six classes was 16.7%, and the maximum color was set to represent 33%.

Although the classification accuracy varied across subjects, general trends can be observed in that: *first*, AEPs elicited by the non-vocoded clean speech (i.e., control; first column of Fig. 6a) were the most distinguishable/classifiable. Here, for illustration purpose, the maximum

color was capped at 33%, whereas the best classification accuracies can often reach 50%. *Second*, although the 22-ch and 4-ch vocoded performance (2nd and 3rd columns) was generally worse than the control AEPs, the difference was relatively small. *Third*, adding background noise to the non-vocoded speech had deteriorating effects on some subjects (4th column), whereas adding background noise to vocoded speech (5th and 6th columns) had clear deteriorating effects on all the subjects.

Figure 6b, c show the grand averages over all the subjects based on the early (0–0.5 s) and late (0.5–1 s) AEPs, respectively. It was clear that the fronto-central area had the highest classification accuracies in both cases. Another area of electrodes that may yield the second-best accuracy was the temporal and/or occipital lobe.

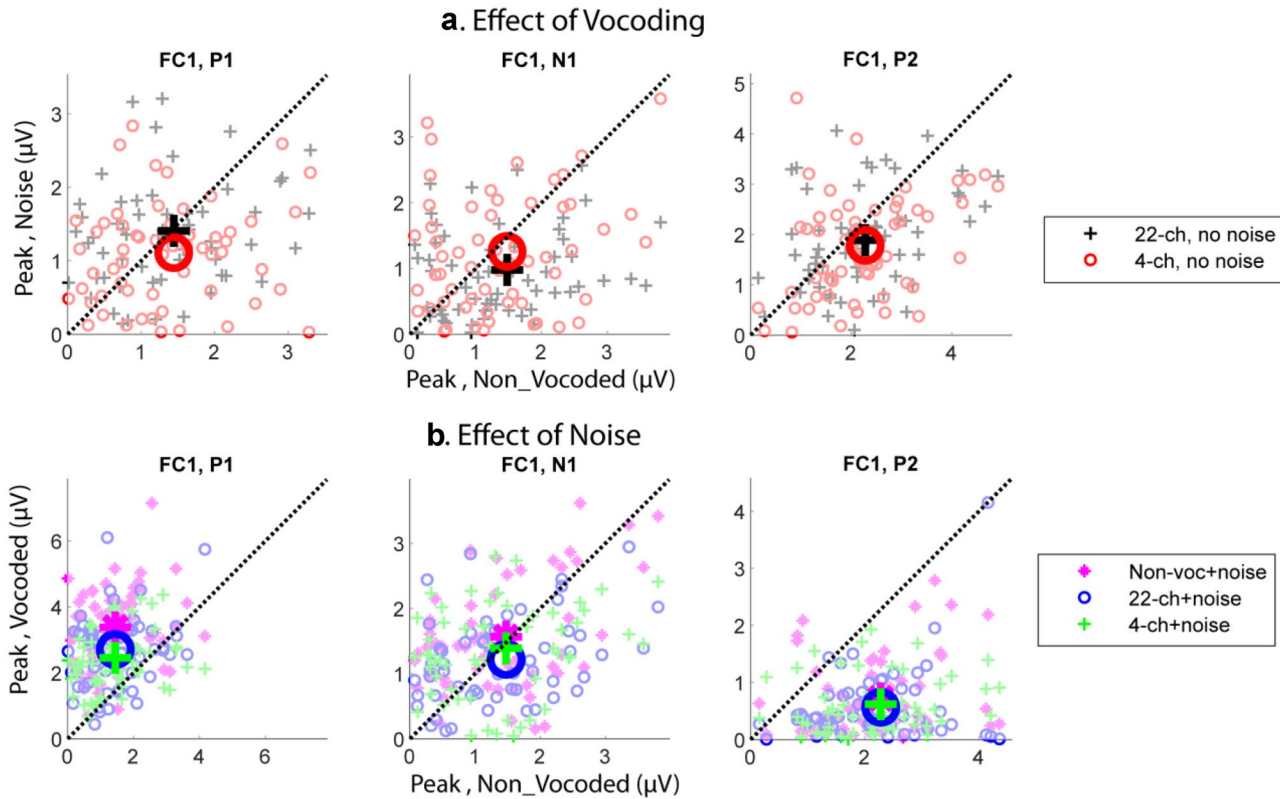


FIG. 4 Effect of vocoding (a) and/or adding background noise (b) on P1, N1, and P2 obtained with electrode FC1. In each scatter plot, the x-axis shows the absolute peak amplitude to the non-vocoded speech without background noise, and the y-axis shows the absolute peak amplitude to speech with vocoding (a) and/or

background noise (b). Each subject had six identical symbols in each sound condition representing the six syllables, and there were a total of 10 subjects. Symbols below the diagonal line would indicate that the peak became smaller with the introduction of vocoding/background noise

When comparing classification performance between the early and late AEPs, some conditions showed significantly or insignificantly better performance with the early portion (t test values in blue near the arrows), whereas other conditions showed better performance with the late portion (t test values in red). In general, the two portions were similar, and we were unable to conclude which half contained better information.

Vocoding and Noise on Cross-condition Correlations

Figures 3, 4, and 5 show that prominent AEP features, such as N1 and P2, can be significantly affected by background noise. Other features of AEPs may be altered too, which can be reflected by performing cross-correlations between AEP waveforms obtained in the control (i.e., clean speech) and any other sound conditions. Figure 7a shows the r_{cc} values averaged over all the subjects. To exclude the clearly noise-induced onset activity around 100–200 ms, we took

the AEP between 0.3 and 1 s for computing all the correlations.

Warm colors indicate high correlation values. For plotting purpose, the maximum r_{cc} was capped at 0.5. In fact, individual r_{cc} values can be as high as 0.9. It seems that the brain response was capable of ignoring the vocoding alone to a great extent. For some subjects (S1, S2, S3, S4, S6, and S7), the majority of electrodes had high r_{cc} values (not shown).

The rightmost three plots of Fig. 6a show the r_{cc} between the clean speech and background noise conditions with or without vocoding. Compared with vocoding-only conditions (the two leftmost plots), r_{cc} was highly significantly reduced. The lowest r_{cc} for each subject was often achieved with the 4-channel vocoding+noise condition (not shown). Using a Holm-Bonferroni correction method, the two leftmost plots were significant ($p < 0.01$).

Similar to the scalp topography of classification (Fig. 6), the best r_{cc} values were again achieved with electrodes Fz, FC1, FC2, and Cz. Electrodes at the occipital lobe yielded high r_{cc} values only for the noise-free conditions.

Figure 7b replots the r_{cc} of Cz as a bar plot. The three noisy r_{cc} 's were significantly lower than the two no-noise r_{cc} 's (z test; the values are specified in the figure).

In general, similar to the classification performance (Fig. 6), the r_{cc} was also decreasing as the speech quality deteriorated. However, a major difference was that the classification plot showed a gradual decrease (Fig. 6b, c), whereas the r_{cc} plot had an abrupt decrease with the introduction of background noise.

One may wonder how those deteriorations in AEP classification and correlation were caused by deteriorations of sound quality. Figure 7c shows the correlations of sound envelopes between the non-vocoded clean speech and other distorted speech. Here, no EEG data was involved; we are only examining the change in the sound waveforms. We did not show the significance values because, due to the large sample size in the sound waveform (e.g., 51,200 points), all the correlation values were significant, and all the pairs of correlations were significantly different.

An interesting observation is that, although a general deteriorating trend existed in the sound envelope (Fig. 7c), it did not represent the trend in either the classification (Fig. 6) or correlation result (Fig. 7b). In fact, it was more similar to the human behavioral result (Fig. 2). This finding implies that the two above analyses we performed on the AEPs most likely did not fully capture human behavior.

EEG Physiology Compared with Human Behavior

As shown above, both AEP classification and r_{cc} showed decreased trends with vocoding and/or background noise. Next, we will examine whether the classification and/or r_{cc} values agreed with the human perception in terms of cross-subject variations. Because the subjects showed a significant amount of errors only with the 4-channel-noise condition (Fig. 2, last bar in each panel), only classification and r_{cc} obtained with this sound condition were examined.

In the example shown in Fig. 8a, we used a scatter plot to compare the r_{cc} value obtained with electrode FC1 and the behavioral data. The statistical values shown in the figure indicate the quality of the linear fit. The linear fit (i.e., correlation) between the AEP classification and behavioral data was highly significant. Recall that r_{cc} here measures the similarity between the clean speech and vocoding/noise AEPs. In other words, those subjects who had high behavioral performance also had high similarities of brain responses using FC1 between clean speech and vocoding/noise speech, indicating better abilities to suppress brain responses to noise with this electrode.

Figure 8b shows this linear fit in terms of a correlation value for all the electrodes, using both AEP classification

and r_{cc} . Only the early portion of the classification seems to explain the human behavior with a small number of electrodes. Overall, electrodes in the fronto-central areas were most consistent in explaining human behavior using early-portion classification or r_{cc} .

Discussion

Effects of Noise Vocoding on EEG Response to Speech

Noise vocoding has been used as a technique to manipulate speech intelligibility, and, to some extent, simulate cochlear implant perception. When there is no background noise, high speech recognition can be obtained with as few as four frequency channels (Shannon et al. 1995). When the listening condition is difficult, more channels are required to maintain intelligibility (Loizou et al. 1999; Friesen et al. 2001; Smith et al. 2002). Similarly, for cochlear implant listeners, speech intelligibility in noise may be effected by the number of spectral channels (Friesen et al. 2001) and characteristics of background noise (Hu and Loizou 2010). Our behavioral result confirmed that more frequency channels are needed for perceiving speech in background noise. Even for 4-channel vocoded speech in quiet, some subjects had problems with certain vowel detections (*/hæd/* and */hɒd/*, Fig. 2b).

With a linear relationship between EEG and sound envelopes varying in between 2 and 8 Hz (Pasley et al. 2012; O'Sullivan et al. 2015), EEG obtained with vocoded speech is expected to more or less resemble the EEG for clean speech. Baltzell et al. (2017) also showed that cross-correlations between speech envelope and brain signals may exist for both clean and noise-vocoded speech.

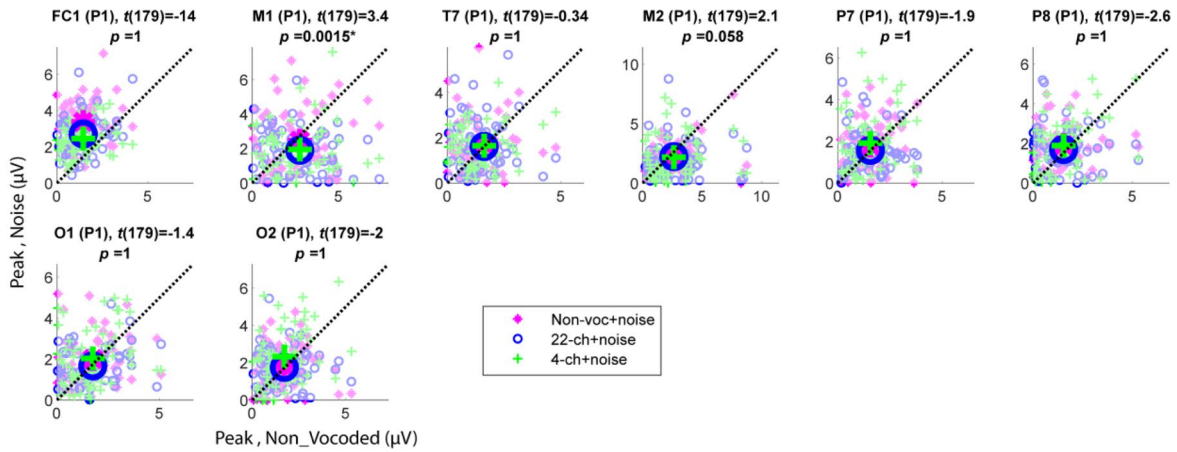
The present study examined the effect of noise vocoding and background noise on the peaks of the AEPs, namely, the P1-N1-P2 complex, as well as two more types of AEP metrics, (1) syllable classification for each sound condition, and (2) cross-condition correlations between the clean speech and distorted speech AEPs.

First, regarding the P1-N1-P2 complex, we did not observe consistent changes in the peak amplitudes (such as Fig. 4a) for vocoded speech without noise. Note that this did not mean there was no change in each individual subject's response. Rather, the effect of vocoding alone on those peaks was not consistent, in contrast to the effect of noise alone.

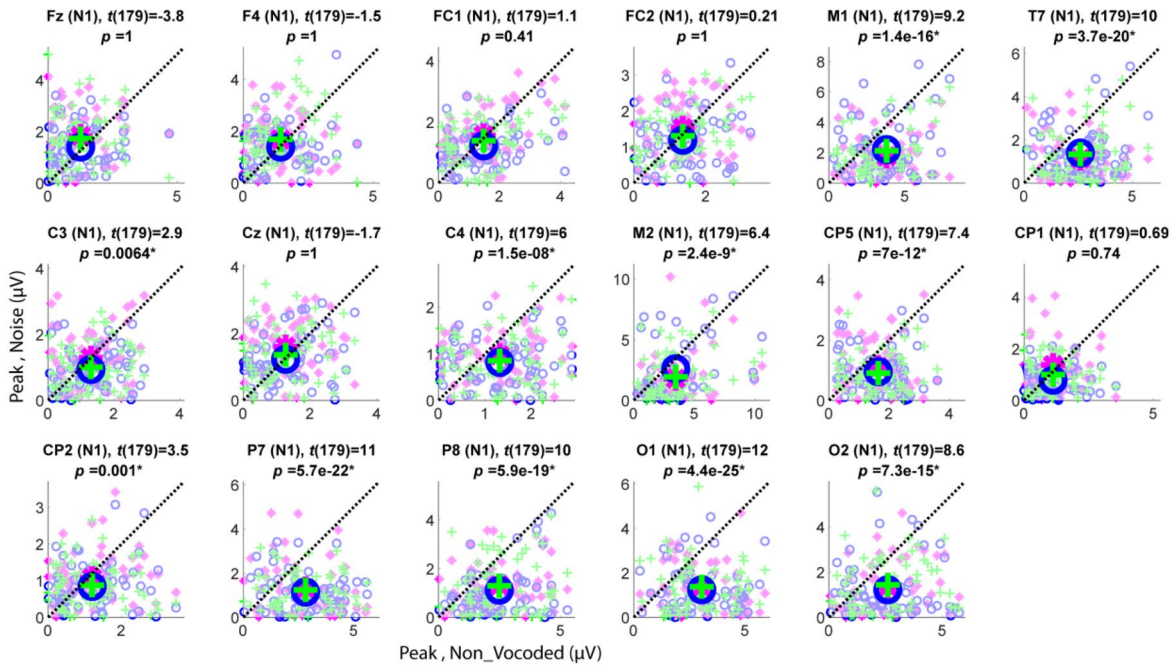
Second, the classification accuracy indicated whether the AEPs of different syllables were consistently distinguishable. Overall, classification accuracy was reduced by noise vocoding, but not as drastically as the background noise.

Third, the correlation metric was a direct measurement of the similarity between AEPs of the clean and

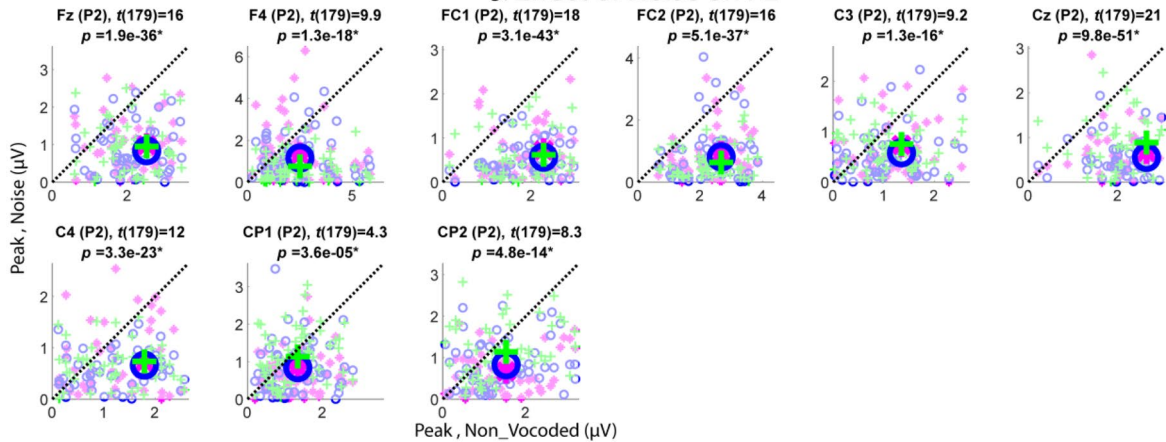
a. Effect of Noise on P1



b. Effect of Noise on N1



c. Effect of Noise on P2



◀ **Fig. 5** Peak analysis on P1, N1, and P2 for all electrodes that showed consistent peaks across subjects according to Table 1. Formats are similar to Fig. 4. Statistics were performed with the one-sided *t* test, with *df*=179; Bonferroni corrections were applied to correct for multiple measurements

vocoded/noisy speech. Most subjects showed significant correlations with both 22- and 4-channel vocoded speech without background noise.

The above observations are reasonable because noise vocoding removes most of the temporal fine structure, while retaining temporal envelope to certain extent in a

few frequency bands. When correlating the sound envelopes of vocoded speech to clean speech (Fig. 7c), we obtained a correlation value of >0.8 for both 22- and 4-channel vocoded speech.

Effects of Background Noise on EEG Response to Speech

Overall, the effect of background noise on the P1-N1-P2 complex and the two other metrics were much more notable than the effect of vocoding alone; however,

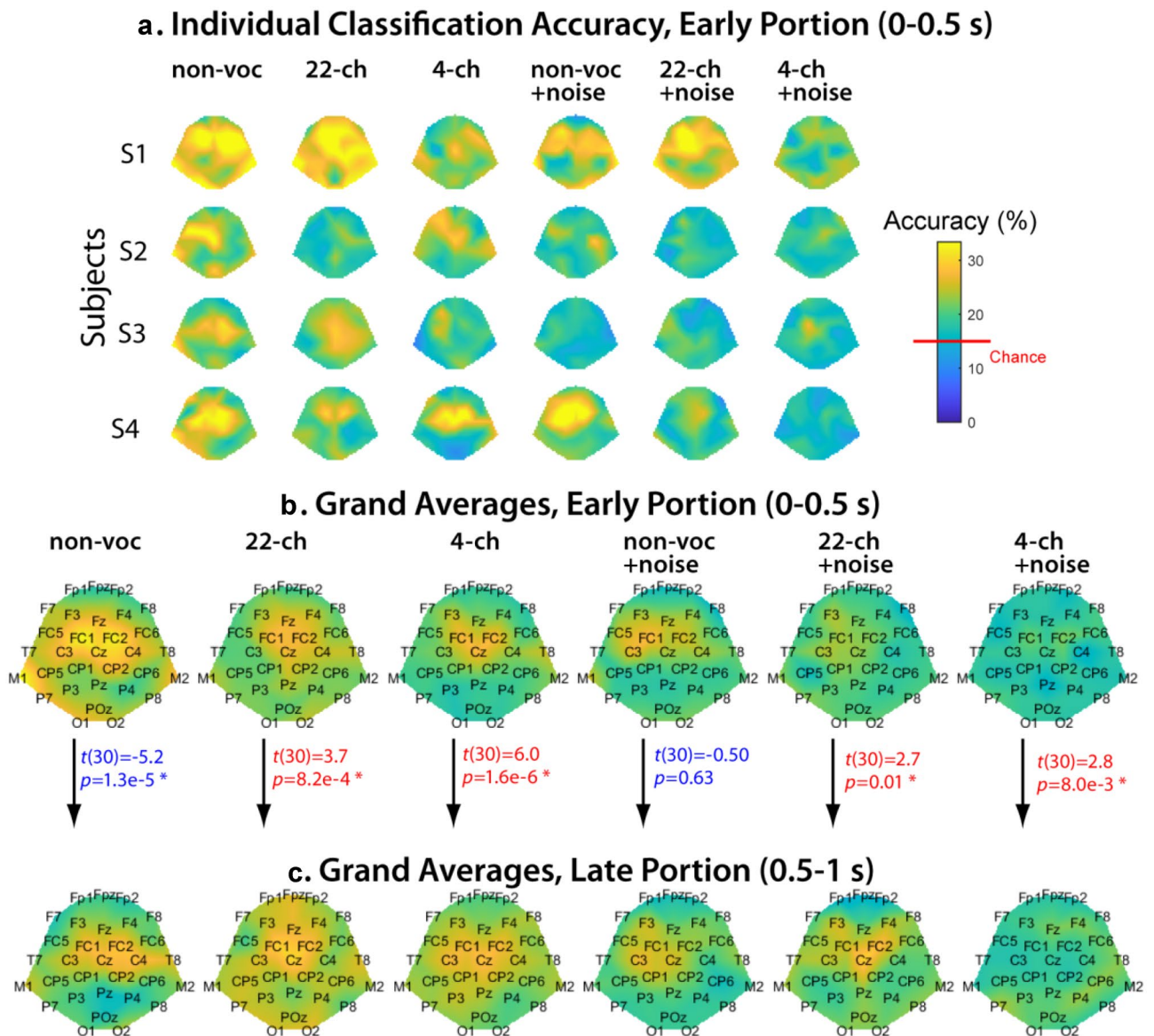
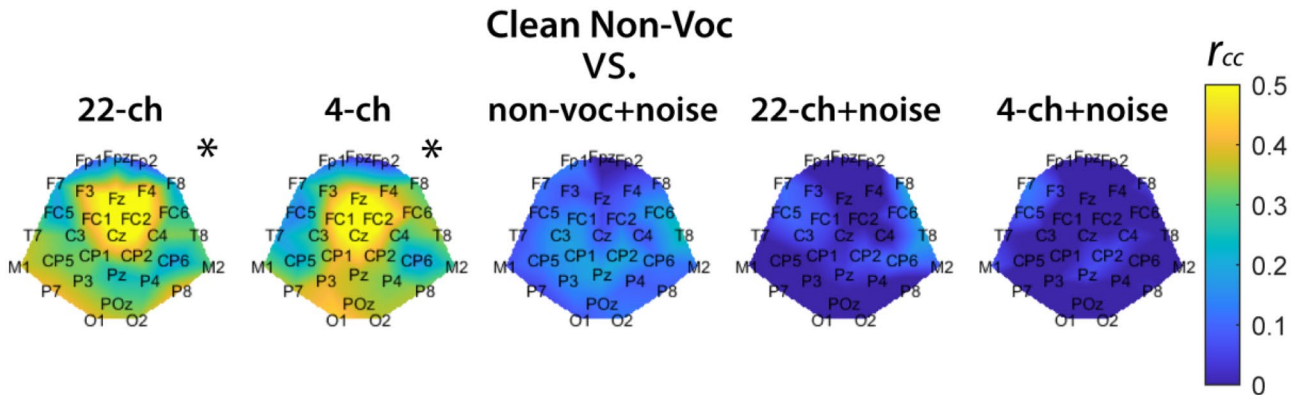


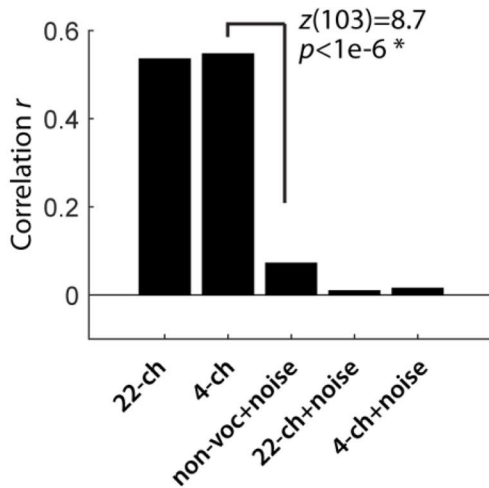
Fig. 6 a Classification accuracies for four representative subjects and sound conditions using the early AEPs (0–0.5 s). Chance performance was 16.7%. b Grand averages over all the subjects for each sound condition using the early AEPs (0–0.5 s). c Grand averages over all the subjects for each sound condition using the late AEPs

(0.5–1 s). The statistical values next to the arrows are two-sided *t* tests (*df*=30) to examine whether early or late AEPs provided significantly better classification accuracy than the early AEPs. A positive *t* value (red) indicates better performance with the two portions. There was not a consistent observation regarding the two portions

a. r_{cc} (Correlations to Clean-Speech AEPs), Grand Averages (0.3-1s)



b. r_{cc}, Cz



c. Sound Envelope Correlations

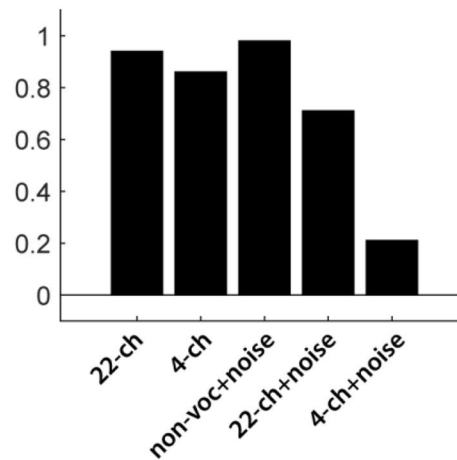


FIG. 7 **a** Cross-condition correlations of AEPs between the non-vocoded clean speech and speech with vocoding and/or noise as grand averages over all the subjects. *The correlations were significant using a Holm-Bonferroni correction method ($p < 0.01$, $df = 103$; exact t values were unavailable due to the nature of performing the Holm-Bonferroni correction). **b** Correlation values obtained with Cz from **a** was replotted as a bar plot. Correlations

of the three conditions with background noise were all significantly lower than the correlations of the two conditions with vocoding only. **c** Correlations of sound envelopes for */hɔd/* between the non-vocoded clean speech and speech with vocoding and/or noise. Here, all the values were significantly different from any other values with the large sample size in the sound waveforms (51,200 points)

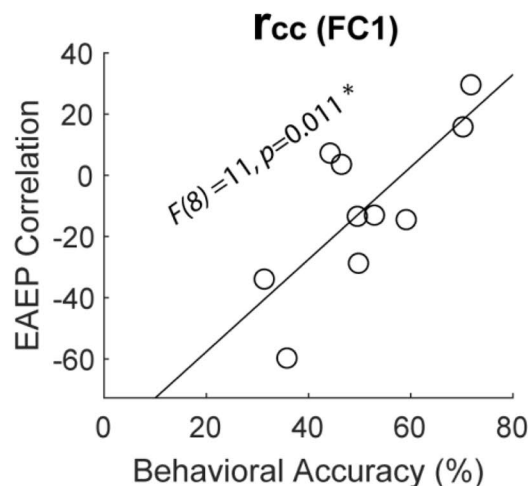
background noise also interacted with the vocoding, jointly impacting the brain responses.

When examining different electrodes, background noise had rarely any consistent effect on P1. It sometimes had a decreasing effect on N1, while it always had a large decreasing effect on P2. This is consistent with the previous observation that (1) N1 and P2 are usually reduced by background noise (Faucette and Stuart 2017) and (2) a late event-related potential (approximately 600 ms following signal onset) is particularly useful for processing degraded speech, or degraded processing of speech (Utianski et al. 2015). It has also been shown that P3 (with a latency similar to our P2) is significantly reduced by the presence of

speech babble noise (Koerner et al. 2016), indicating that our observation is not limited to white noise.

Both the classification and r_{cc} metrics were affected by background noise more than noise vocoding. When background noise was added to vocoded speech, both metrics showed further reductions in performance. However, there existed a very clear distinction between the behaviors of the two metrics. With background noise alone and no vocoding, the classification metric based on grand averages only decreased slightly (Fig. 6), whereas the correlation metric showed a highly significant decrease (Fig. 7a, b). This decrease of correlation was not a noise onset effect, because we excluded the first 0.3 s from the computation of the

a. Example of AEP Correlation vs. Behavior (4-ch + Noise)



b. AEP Analysis vs. Behavior (4-ch + Noise), All Electrodes

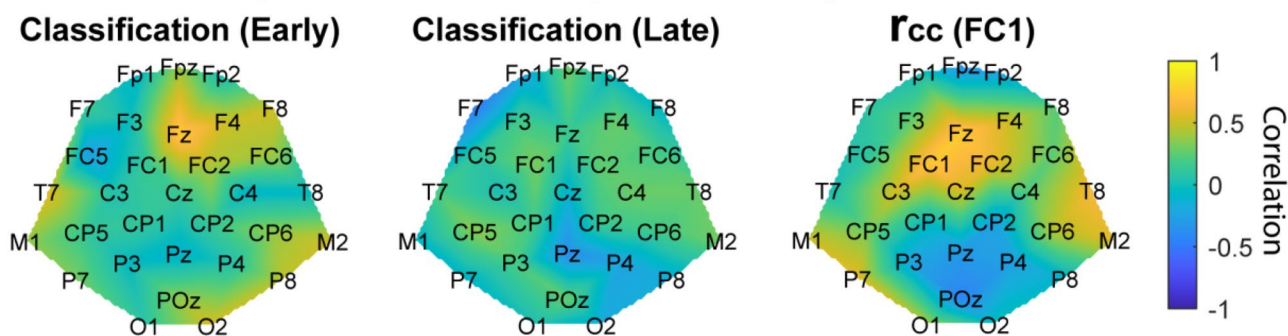


Fig. 8 Explaining human behavioral data using AEP metrics, which are either the classification accuracy or r_{cc} . Here, only human performance with the 4-channel vocoding and noise condition was examined. **a** An example scatter plot of the behavioral accuracy

and AEP correlation (r_{cc}) obtained with electrode FC1. **b** The same as in **a** except that all the electrodes were plotted for both AEP classification and r_{cc} . The color represents the correlation between the AEP analysis and the human behavior

correlation value. Instead, it was an ongoing suppression of speech-related features.

This decrease was also not an effect of deteriorated speech envelope, because the sound envelope with background noise remained highly correlated to the speech envelope (Fig. 7c, third bar). Overall, the trend in the sound envelope correlation (Fig. 7c) was most similar to the behavioral data (Fig. 2). Next was the classification metric (Fig. 6) that decreased gradually. The r_{cc} metric (Fig. 7) was the least similar to the behavioral result in terms of this decreasing trend. This finding implies that neither the classification nor the correlation method fully captures human behavior.

For speech encoding, two brain areas, the lateral posterior temporal and fronto-central electrode sites, are especially interesting. The two areas have been shown to reliably display syllable-specific waveforms (Wagner et al. 2016). In particular, Cz is a commonly used electrode for speech perception and syllable

classification (Brandmeyer et al. 2013; Jafarpisheh et al. 2016; Khalighinejad et al. 2017; Steinmetzger and Rosen 2017). In our study, Cz, FC1, FC2, and Fz generally yielded the best classification and correlation performance, with a slight preference to the left side (FC1).

EEG Metrics and Human Behavior

In the present study, we created a difficult listening condition using the 4-channel vocoded speech presented in background noise. A significant number of behavioral errors were obtained. Subjects also showed variable detection performance with this condition (Fig. 2a). We compared this behavioral variability with the two EEG metrics to examine if they can explain the cross-subject variation.

We found a significant correlation between the behavioral performance and cross-condition AEP correlation

(r_{cc}). In other words, subjects who had high similarities in their electrophysiology between the 4-channel vocoded noisy AEPs and clean speech AEPs also performed better behaviorally. This is reasonable because, if the subject's cortical signals can actively suppress noise, fewer behavioral errors should have been made.

This study did not directly examine cortical entrainment. However, the assumption was that the AEPs we obtained were more or less determined by speech envelope. In particular, the frequency range of 2–8 Hz in cortical responses is found to be linearly correlated to speech envelope (Pasley et al. 2012; Zion Golumbic et al. 2013; O'Sullivan et al. 2015). Slow and moderate temporal fluctuations in speech may be linearly reconstructed from cortical signals even in certain background noise (Ding et al. 2014; Puvvada et al. 2017), whereas fast temporal fluctuations require nonlinear transformations. The correlation metric may provide answer to the question—when sound quality changes, especially when speech is degraded by vocoding or background noise, can the brain resist those non-preferred changes and still respond as if clean speech were being presented? In this sense, a high correlation between human behavior and r_{cc} is reasonable.

When the classification metric was compared with behavioral data, only the early portion (0–0.5 s) of AEPs showed high correlations (Fig. 8b). This is surprising because the classification performance was comparable for the early and late AEPs (Fig. 6b, c). Note that we only had 10 subjects for the linear fit. Future studies will need a larger subject population to validate these relationships.

In summary, the present study examined the effect of noise vocoding and background noise on the AEP peaks (i.e., P1, N1, and P2) and two other metrics (i.e., classification and correlation). Background noise generally had larger deteriorating effects than noise vocoding alone, but it also interacted with vocoding in generating the worst human and EEG performance. Last, noise suppression in the EEG signals did seem to correlate with human behavior.

ACKNOWLEDGEMENTS

We thank R. Shannon for sharing the consonant and vowel audio files. We also thank the two reviewers for bring up multiple suggestions that greatly improved our analyses.

REFERENCES

BALAKRISHNAMA S, GANAPATHIRAJU A, PICONE J (1999) Linear discriminant analysis for signal processing problems. In: Southeastcon '99. Proceedings. IEEE, pp 78–81: Southeastcon '99. Proceedings. IEEE

BALTZELL LS, SRINIVASAN R, RICHARDS VM (2017) The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *J Neurophysiol* 118:3144–3151

BEEKHUIZEN B, BOD R, ZUIDEMA W (2013) Three design principles of language: the search for parsimony in redundancy. *Lang Speech* 56:265–290

BRANDMEYER A, FARQUHAR JD, MCQUEEN JM, DESAIN PW (2013) Decoding speech perception by native and non-native speakers using single-trial electrophysiological data. *PLoS One* 8:e68261

DIGESER FM, WOHLBEREDT T, HOPPE U (2009) Contribution of spectrotemporal features on auditory event-related potentials elicited by consonant-vowel syllables. *Ear Hear* 30:704–712

DING N, CHATTERJEE M, SIMON JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88:41–46

EISENBERG LS, SHANNON RV, MARTINEZ AS, WYGONSKI J, BOOTHROYD A (2000) Speech recognition with reduced spectral cues as a function of age. *J Acoust Soc Am* 107:2704–2710

FAUCETTE SP, STUART A (2017) Evidence of a speech evoked electrophysiological release from masking in noise. *J Acoust Soc Am* 142:EL218

FRIESEN LM, SHANNON RV, BASKENT D, WANG X (2001) Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am* 110:1150–1163

FRIESEN LM, TREMBLAY KL, ROHILA N, WRIGHT RA, SHANNON RV, BASKENT D, RUBINSTEIN JT (2009) Evoked cortical activity and speech recognition as a function of the number of simulated cochlear implant channels. *Clin Neurophysiol* 120:776–782

GLASBERG BR, MOORE BC (1990) Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47:103–138

HILLENBRAND J, GETTY LA, CLARK MJ, WHEELER K (1995) Acoustic characteristics of American English vowels. *J Acoust Soc Am* 97:3099–3111

HU Y, LOIZOU PC (2010) Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *J Acoust Soc Am* 127:3689–3695

JAFARISHEH AS, JAFARI AH, ABOLHASSANI M, FARHADI M, SADJEDI H, POURBAKHT A, SHIRZHIYAN Z (2016) Nonlinear feature extraction for objective classification of complex auditory brainstem responses to diotic perceptually critical consonant-vowel syllables. *Auris Nasus Larynx* 43:37–44

KHALIGHINEJAD B, CRUZATTO DA SILVA G, MESGARANI N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. *J Neurosci* 37:2176–2185

KIM J, LEE SK, LEE B (2014) EEG classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition. *J Neural Eng* 11:036010

KOERNER TK, ZHANG Y, NELSON PB, WANG B, ZOU H (2016) Neural indices of phonemic discrimination and sentence-level speech intelligibility in quiet and noise: a mismatch negativity study. *Hear Res* 339:40–49

LOCK RH, LOCK PF, MORGAN KL, LOCK EF, LOCK DF (2017) *Statistics: Unlocking the Power of Data*, 2nd edn. Wiley, NJ

LOIZOU PC, DORMAN M, TU Z (1999) On the number of channels needed to understand speech. *J Acoust Soc Am* 106:2097–2103

MUGLER EM, PATTON JL, FLINT RD, WRIGHT ZA, SCHUELE SU, ROSENOW J, SHIH JJ, KRUSIENSKI DJ, SLUTZKY MW (2014) Direct classification of all American English phonemes using signals from functional speech motor cortex. *J Neural Eng* 11:035015

O'SULLIVAN JA, POWER AJ, MESGARANI N, RAJARAM S, FOXE JJ, SHINN-CUNNINGHAM BG, SLANEY M, SHAMMA SA, LALOR EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb Cortex* 25:1697–1706

PASLEY BN, DAVID SV, MESGARANI N, FLINKER A, SHAMMA SA, CRONE NE, KNIGHT RT, CHANG EF (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251

- PEELLE JE, GROSS J, DAVIS MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23:1378–1387
- PUVVADA KC, VILLAFANE-DELGADO M, BRODBECK C, SIMON JZ (2017) Neural Coding of Noisy and Reverberant Speech in Human Auditory Cortex. *bioRxiv*:229153
- SHANNON RV, FU QJ, GALVIN J, 3rd (2004) The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol Suppl*: 50–54
- SHANNON RV, ZENG FG, KAMATH V, WYGONSKI J, EKELID M (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304
- SHANNON RV, JENSVOLD A, PADILLA M, ROBERT ME, WANG X (1999) Consonant recordings for speech testing. *J Acoust Soc Am* 106:L71-74
- SMITH ZM, DELGUTTE B, OXENHAM AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416:87–90
- STEINMETZGER K, ROSEN S (2017) Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia* 95:173–181
- STILP CE (2011) The redundancy of phonemes in sentential context. *J Acoust Soc Am* 130:EL323–328
- SWINK S, STUART A (2012) Auditory long latency responses to tonal and speech stimuli. *J Speech Lang Hear Res* 55:447–459
- UTIANSKI RL, CAVINESS JN, LISS JM (2015) Cortical characterization of the perception of intelligible and unintelligible speech measured via high-density electroencephalography. *Brain Lang* 140:49–54
- WAGNER M, ROYCHOUDHURY A, CAMPANELLI L, SHAFER VL, MARTIN B, STEINSCHNEIDER M (2016) Representation of spectro-temporal features of spoken words within the P1-N1-P2 and T-complex of the auditory evoked potentials (AEP). *Neurosci Lett* 614:119–126
- WEISZ N, OBLESER J (2014) Synchronisation signatures in the listening brain: a perspective from non-invasive neuroelectrophysiology. *Hear Res* 307:16–28
- WON JH, CLINARD CG, KWON S, DASIKA VK, NIE K, DRENNAN WR, TREMBLAY KL, RUBINSTEIN JT (2011) Relationship between behavioral and physiological spectral-ripple discrimination. *J Assoc Res Otolaryngol* 12:375–393
- ZHOU W, GOTMAN J (2009) Automatic removal of eye movement artifacts from the EEG using ICA and the dipole model. *Prog Nat Sci* 19:1165–1170
- ZION GOLUMBIC EM, DING N, BICKEL S, LAKATOS P, SCHEVON CA, MCKHANN GM, GOODMAN RR, EMERSON R, MEHTA AD, SIMON JZ, POEPEL D, SCHROEDER CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron* 77:980–991

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.