Research Article

Effects of Spectral Degradation on Attentional Modulation of Cortical Auditory Responses to Continuous Speech

YING-YEE KONG,^{1,2} ALA SOMAROWTHU,² AND NAI DING³

¹*Department of Communication Sciences and Disorders, Northeastern University, Boston, MA 02115, USA*

²*Department of Bioengineering, Northeastern University, Boston, MA 02115, USA*

³*College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Zhejiang, China*

Received: 26 October 2014; Accepted: 24 August 2015; Online publication: 11 September 2015

ABSTRACT

This study investigates the effect of spectral degradation on cortical speech encoding in complex auditory scenes. Young normal-hearing listeners were simultaneously presented with two speech streams and were instructed to attend to only one of them. The speech mixtures were subjected to noise-channel vocoding to preserve the temporal envelope and degrade the spectral information of speech. Each subject was tested with five spectral resolution conditions (unprocessed speech, 64-, 32-, 16-, and 8-channel vocoder conditions) and two target-to-masker ratio (TMR) conditions (3 and 0 dB). Ongoing electroencephalographic (EEG) responses and speech comprehension were measured in each spectral and TMR condition for each subject. Neural tracking of each speech stream was characterized by cross-correlating the EEG responses with the envelope of each of the simultaneous speech streams at different time lags. Results showed that spectral degradation and TMR both significantly influenced how top-down attention modulated the EEG responses to the attended and unattended speech. That is, the EEG responses to the attended and unattended speech streams differed more for the higher (unprocessed, 64 ch, and 32 ch) than the lower (16 and 8 ch) spectral resolution conditions, as well as for the higher (3 dB) than the lower TMR (0 dB) condition. The magnitude of differential neural modulation responses to the attended and unattended speech streams significantly

correlated with speech comprehension scores. These results suggest that severe spectral degradation and low TMR hinder speech stream segregation, making it difficult to employ top-down attention to differentially process different speech streams.

Keywords: cortical entrainment, cochlear implant, spectral degradation, sound segregation, event-related potentials, auditory scene analysis

INTRODUCTION

Successful speech comprehension in a complex background requires listeners to first segregate different sound sources and then apply top-down attention to select one source for further processing. Listeners' ability to selectively attend, therefore, depends on their ability to analyze the complex auditory scene and to form perceptual auditory objects (Shinn-Cunningham 2008). Previous studies have demonstrated that auditory stream segregation and object formation are influenced by cues varying from low-level spectro-temporal acoustic features (e.g., harmonics) to higher-level perceptual features (pitch, timbre, spatial location) and linguistic knowledge (see Bregman 1990; Shinn-Cunningham and Best 2008 for review). Acoustically, sounds that have common onset or amplitude modulation (e.g., Culling and Summerfield 1995; Best et al. 2007) and sounds that are harmonically related tend to group together (e.g., Bregman 1990; Darwin and Carlyon 1995). Higher-

Correspondence to: Ying-Yee Kong · Department of Communication Sciences and Disorders · Northeastern University · Boston, MA 02115, USA. Telephone: +1 (617) 373-3704; email: yykong@neu.edu

order perceptual features, such as pitch, timbre, and spatial location, also facilitate the formation of an auditory object (e.g., Culling and Darwin 1993; Darwin and Hukin 2000a, b; Kidd et al. 2005) and direct listeners' attention to the object of interest. The neurally segregated auditory objects are then selectively modulated based on top-down attention (e.g., Buschman and Miller 2007; Elhilali et al. 2007; Fritz et al. 2007).

Recent electrophysiological studies on normal-hearing listeners provided further evidence for object-based processing during speech stream segregation (e.g., Ding and Simon 2012a; Kerlin et al. 2010; Mesgarani and Chang 2012), in which attention enhances cortical responses to the attended speech stream and reduces the responses to the ignored speech stream (e.g., Horton et al. 2013, 2014; Kong et al. 2014; O'Sullivan et al. 2014). Ding and Simon (2012a) varied the attended and background speakers separately over a wide range of intensity and found that the neural representation of the attended speech did not adapt to the intensity of the background speaker, which strongly suggests an object-level intensity gain control. The attentional effect was observed at both early (<100 ms) (Ding and Simon 2012a; Kong et al. 2014) and later (>150 ms) (Horton et al. 2013; O'Sullivan et al. 2014; Kong et al. 2014) processing stages. By cross-correlating the electroencephalography (EEG) responses with the envelopes of individual speakers in a two-speaker speech mixture, Kong et al. (2014) showed salient peaks around 100 and 200 ms in the cross-correlation function for the attended speech but a delayed and much attenuated peak at around 200 ms for the unattended speech. In sum, these results suggest that when different speech streams are perceptually segregated, neural responses differentially track different auditory objects with the attended signal being amplified and the unattended signal(s) suppressed (Horton et al. 2013; Kong et al. 2014).

To date, most research has focused on listening situations where the listeners can easily segregate simultaneous sound streams and maintain attention on one using pitch or spatial cues. However, it is yet to be determined how attention modulates cortical responses when the simultaneous speech streams cannot be easily segregated. For example, many of the auditory cues that enable object formation are not well encoded in individuals with hearing impairment or with a cochlear implant due to reduced temporal and spectral resolution, as well as reduced acuity for spatial location (see Moore 2007 for review). The perception of the pitch and spectral cues requires precise coding of the spectral and timing information. For example, spectral degradation negatively affects listeners' ability to segregate competing speech signals

based on pitch harmonic cues (e.g., Stickney et al. 2004; Qin and Oxenham 2005; see Oxenham 2008 for review). If selective attention operates on auditory objects, the failure of auditory object formation due to reduced sound segregation would hinder a selective neural representation of the sound sources of interest.

The current study employed the noise-channel vocoding technique to systematically investigate the effect of spectral degradation on cortical responses to a speech mixture. By manipulating the number of frequency channels used in the vocoding processing, we examined the extent to which attentional modulation of neural responses is affected by the severity of spectral degradation of speech. Ding et al. (2014) examined the neural responses to vocoder speech embedded in stationary broadband noise and found that delta band neural response enhanced with increasing spectral resolution and this enhancement correlated with speech comprehension scores. Rimmele et al. (2015) further demonstrated that spectral resolution influences not only neural tracking of speech but also differentially influences the neural tracking of attended and unattended speech streams, in the case when the attended and unattended speech streams had different spectral resolutions. Different from these studies, we (1) systematically examined the effect of spectral degradation on selective attention in a competing background by varying the number of vocoder channels and (2) investigated the relationship between neural modulation responses and speech comprehension ability.

Figure 1 displays spectrograms of an unprocessed speech signal (left) and the same speech stimuli processed with a 64 (middle) and 8 (right) channel noise vocoder. The harmonic structure and the spectral details of the speech signal are clearly observed in the unprocessed speech but decrease with decreasing number of channels in the vocoder speech. The red line that is superimposed on each panel represents the intensity changes of the speech signal over time (i.e., temporal envelope of speech). As seen in this figure, the slowly varying temporal envelopes of speech are similar between different vocoder conditions, and they are also similar to the original speech signal. Selective auditory attention to the target speech, if successful, could lead to differential neural modulation to the temporal envelopes of individual vocoder speech streams, as we have observed with unprocessed speech (Kong et al. 2014). We hypothesized that a more severe spectral degradation will lead to a more frequent failure of speech segregation. Furthermore, a failure of sound segregation or auditory object formation will prohibit the brain from selectively modulating the neural responses to individual speech streams.

Fig. 1. Spectrograms of unprocessed (*left*), 64-channel vocoder (*middle*), and 8-channel vocoder (*right*) speech. The *red line* on each panel represents the temporal envelope of the broadband speech.

METHODS

Participants

Eight healthy normal-hearing, right-handed (Oldfield 1971) adult native speakers of American English between the ages of 20 and 28 years participated in the study. Hearing screening at 20 dB HL across frequencies (250 to 8000 Hz) for each ear was performed in a double-walled sound-isolated booth on all subjects to confirm normal hearing status. Questions related to history of speech and language impairments, overall health, use of medications, handedness, and history of tinnitus were included in a questionnaire, which each subject needed to fill out to further determine their eligibility. Those who had a history of speech and language disorders, tinnitus, or consumption of medications that could affect brain functions were excluded from the study. This study was conducted according to the protocols approved by the Institutional Review Board of Northeastern University. Subjects received compensation for their participation. Written informed consent was obtained prior to the experiment.

Stimuli

Auditory stimuli were continuous speech extracted from two chapters in a public domain children's book, "A Child's History of England" by Charles Dickens (<http://librivox.org/a-childs-history-of-england-by-charles-dickens/>). One chapter (chapter 22) was read by a male speaker, and the other (chapter 35) was read by a female speaker. The audio files were recorded at a sampling rate of 22.05 kHz. To maintain a continuous flow of speech, all silent intervals in the recordings longer than 300 ms were shortened to

300 ms. Each extracted passage was divided into segments with durations of approximately 1 min. The actual length of the segment varied slightly so that the end of the audio file included a complete sentence. We then normalized all of the 1-min speech segments to have equal root-mean-square (RMS) amplitude. Speech mixtures were constructed by mixing two speakers digitally. In each mixture, one speaker began 1 s after the other speaker, and both speakers ended at the same time. The RMS level of one speaker was fixed at 65 dB in the mixture, while the other speaker that had a delayed start was either the same or 3 dB weaker. This resulted in two target-to-masker ratio (TMR) conditions of 0 and 3 dB, respectively. The stimuli were presented diotically using insert earphones. All experiments were conducted in a double-walled sound-isolated booth.

The speech mixtures were then subjected to noise-channel vocoding to simulate cochlear-implant sound processing, which preserves the temporal envelope of speech but degrades the spectral and temporal fine structure. The system used for channel-vocoding processing was similar to that described by Shannon et al. (1995). In this system, broadband speech (80–8000 Hz) was band-pass filtered into a number of logarithmically spaced frequency bands using the Greenwood map (Greenwood 1990). The amplitude envelope of the signal was extracted from each band using the Hilbert transform followed by LP filtering with a 300-Hz cutoff frequency. The envelope extracted from each frequency band was used to modulate white noise, which was then filtered by the same band-pass filter used to generate the frequency band in the analysis stage. All bands were then summed to produce the final vocoder stimuli. All vocoder stimuli were scaled to match the RMS of the unprocessed

stimuli. The number of frequency bands used in this study varied from 8, 16, 32, and 64 bands. The range of vocoder channel numbers was chosen to cover a wide range of speech perception performance with a single-talker masker at the TMRs tested (e.g., Stickney et al. 2004).

Test Conditions and Procedures

Prior to the main experiments, each subject was presented with 150 repetitions of a short (100-ms with 10-ms on- and off-ramps) 1000-Hz tone pip to elicit the auditory N1 response. After confirming that N1 response was present for the subject, (s)he proceeded with the experiment. As part of a larger study on neural responses to distorted speech, the current paper presents findings to contrast neural responses to the unprocessed and vocoder speech. Each subject was tested with five spectral conditions including one unprocessed (unproc) speech condition and four vocoder speech conditions (64, 32, 16, and 8 ch).

For each spectral condition, two randomly chosen 1-min speech mixtures for each TMR condition were presented to the subjects. Across subjects, no one single 1-min recording was used more than once for the same spectral and TMR condition. Also, no speech mixture was repeated across test conditions within subject. Each subject was presented with four testing blocks (2 speech mixtures $\times$ 2 target speakers) per spectral condition per TMR. Each block consisted of one speech mixture that was repeated 10 times. Specifically, two of the four blocks had the same speech mixture, but subjects were asked to pay attention to different speakers in the mixture for different blocks. Block 1 and block 2 presented different speech mixtures, and the subjects were asked to focus on the male speaker in block 1 and the female speaker in block 2. In block 3 and block 4, the speech mixtures were the same as those played in block 1 and block 2 respectively, but the subjects switched their attention to the other speaker.

At the beginning of each trial, a written text cue (the word “male” or “female”) was displayed on a computer screen to indicate which speaker the subject should pay attention to in the mixture. The cue lasted for 1 s and was then replaced by a crosshair “+.” Subjects were asked to maintain visual fixation on the crosshair for the duration of the trial. For each mixture, the unattended speaker started 1 s after the attended speaker to help the subject attend to the correct speaker as cued. Blocks 1 and 2 for each of the two TMR conditions were tested in one test session. Blocks 3 and 4 were tested in a different session conducted on a different day. The whole experiment was divided into one training session and six test sessions administrated on different days. Each session

consisted of two blocks (blocks 1 and 2, or blocks 3 and 4) of testing for one or two spectral conditions at each TMR. For example, test session 1 consisted of blocks 1 and 2 for unproc and 64-ch conditions at 3- and 0-dB TMR; session 2 consisted of blocks 3 and 4 for the same spectral and TMR conditions.

The subjects were familiarized with the vocoder speech before the experiment. The familiarization section started with the easiest condition of unprocessed stimuli and then progressed to the hardest 8-ch vocoder condition. For each spectral condition, subjects were presented with two 1-min duration speech mixtures at 20-, 10-, 3-, and 0-dB TMR, in the order of decreasing TMR. The 20- and 10-dB conditions were only presented in the familiarization section but not tested during data collection. For each mixture, subjects were asked to pay attention to one speaker first and then switched to the other speaker when it was played again. Thus, there were 4 min of training speech materials per spectral condition per TMR (5 spectral conditions $\times$ 4 TMRs). The training session lasted for approximately 2 h. For the subsequent test sessions, each subject received another 4-min practice listening to each of the spectral and TMR condition tested on that day before the experiment with EEG recording.¹

For each subject, the *unproc* condition was tested first, followed by spectral conditions. The order of the spectral conditions was random. Within each spectral condition, the higher TMR (3 dB) was always tested first followed by the lower TMR condition (0 dB). The fixed order of TMR from easier (3-dB TMR) to harder (0-dB TMR) within each test session allows listeners more time to familiarize with vocoder speech for the more difficult listening condition. There was a 15-min break between TMR conditions. Table 1 summarizes the spectral and TMR test conditions for different test sessions (note that different sessions were administrated on different days).

Subjects were asked to answer six open-ended or multiple-choice questions for the attended speech after each block, resulting a total of 24 questions per

¹ The amount of training subjects received for each spectral condition should be sufficient. Our previous work on vocoder speech perception established that the acclimatization process to vocoder speech in quiet was similar across young normal-hearing listeners (e.g., Kong et al. 2015), where performance plateaued after 30 short sentences. Once trained, these listeners' ability to understand vocoder speech in quiet lasted after the training session into later days. As for training on vocoder speech perception in a competing talker condition, we decided to provide 4 $\times$ the amount of training compared to the quiet condition to balance between sufficient training time and the length of the test session (i.e., about 2 h per session). The amount of training included 12 min per spectral condition per TMR (4 min during the training session, 4 min during the first test session of the test condition, 4 min during the second test session of the same test condition).

Table 1

Summary of test conditions and test sessions

Session num ^a	Unprocessed	Spectral cond. 1 ^b	Spectral cond. 2	Spectral cond. 3	Spectral cond. 4
1	Blocks 1–2 per TMR	Blocks 1–2 per TMR			
2	Blocks 3–4 per TMR ^c	Blocks 3–4 per TMR			
3			Blocks 1–2 per TMR	Blocks 1–2 per TMR	
4			Blocks 3–4 per TMR	Blocks 3–4 per TMR	
5					Blocks 1–2 per TMR
6					Blocks 3–4 per TMR

Note that each subject participated in one training session that lasted from 2–3 hours for familiarization of vocoder speech prior to testing

^aThree-decibel TMR was tested before 0-dB TMR in each session

^bNote that the order of testing of the spectral conditions (i.e., number of vocoder channels) was randomized across subjects

^cBlocks (1–2) and blocks (3–4) contained the same speech mixtures, but the speaker in the mixture to whom the listener attended to was switched from one session to another

spectral per TMR condition. Approximately half of the 1-min recordings contained only open-ended questions (six questions), and the rest contained one multiple-choice question and five open-ended questions. We carefully constructed the questions such that they tested comprehension of the passage (i.e., understanding of the main story line) instead of memorizing numbers or names, etc. Subjects paced themselves during the experiment. After each presentation of the speech mixture, subjects clicked on the “continue” button displayed on the computer screen to receive the next presentation of the mixture. After answering the comprehension questions in each block, subjects were asked to take a break that lasted for 10–15 min.

Data Recording and Analyses

EEG signals were recorded using a 16-channel BrainVision actiCHamp system in a double-walled sound-isolated booth, at a sampling rate of 500 Hz. The placement of the electrodes followed the International 10/20 system (Jasper 1958). The recording electrode locations were Fp1, Fp2, F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, O1, Oz, O2, T7, and T8. A reference electrode was placed on the left mastoid.

Due to the fact that the speech stimuli were not exactly equal in length (i.e., 1 min), subsequent analyses were performed on the first 50 s of the ongoing neural responses to the speech signal. In addition, we excluded the first second of the EEG recordings after the mixture onset to minimize the effect of the onset response. For each subject, the EEG signals were then filtered between 2 and 40 Hz, using a fifth-order Butterworth filter. To minimize filter delay, zero phase filtering was performed using the *filtfilt* command in MATLAB. The stimulus phase-locked response component was extracted using denoising source separation (DSS; de Cheveigne and Simon 2008). This technique computes response components that show the highest reliability across

trials. The DSS components are sorted based on their reliability across trials with repeated stimulus. The first DSS component, which had been previously shown to be effective in capturing envelope-tracking neural activity measured by EEG (Kong et al. 2014) and MEG (Ding and Simon 2012b), was further analyzed. The topography of the first DSS component resembles the topography of the auditory N1 response, showing maximal responses at the Cz, C3, and C4 channels (see Fig. 2). This pattern of results is similar to that reported in previous EEG studies on speech envelope tracking (Lalor et al. 2009; Power et al. 2012). We have also analyzed the data from channel Cz alone, and the results are highly consistent with the DSS results.

The temporal envelope of each speech stream in the mixture was extracted using the Hilbert transform. To match the sampling rate and the bandwidth of the speech signal and the EEG signals, the speech envelopes were resampled to a sampling rate of 500 Hz, followed by bandpass filtering between 2 and 40 Hz. The same fifth-order Butterworth filter and zero-phase filtering technique described previously for the EEG signal were used for the speech signals. This further enables the analysis to account for any filter delay.

From each subject and each test condition, the cross-correlation between the speech envelope and the EEG responses was calculated over the entire speech segment duration, i.e., 49 s, for time lags varying from –200 to 600 ms at every 2-ms intervals. The cross-correlation functions were separately computed between the EEG signals and the temporal envelopes of the unprocessed attended and unattended speech. The computation of cross-correlation is described in Eq. (1):

$$C(s, r, dt) = \frac{1}{N-dt} \sum_{i=1}^{N-dt} \frac{(s_i - \bar{s})(r_{i+dt} - \bar{r})}{\sigma_r \sigma_s} \quad (1)$$

where N is the length of the signal (i.e., 49 s), and $\bar{s}$, $\bar{r}$,

Fig. 2. Topography of the spatial filter coefficients for the first DSS component averaged across subjects.

σ_s , and σ_r are the mean and standard deviations of the speech signal and neural response, respectively. For example, for speech signal presented at time t , the cross-correlation between the speech envelope and the neural response at t will yield a correlation coefficient value at time delay of 0 ms. The cross-correlation of the speech envelope at time t and the neural response at $t+dt$ will yield the correlation coefficient at a time delay (i.e., time lag) of dt .

For each subject and test condition, a neural modulation index (NMI) was calculated. NMI was computed as the RMS of the difference in the cross-correlation values between the attended and unattended speech over time lags. Thus, NMI is a way to quantify how much attention modulates the cross-correlation between EEG responses and individual speech streams.

The group level cross-correlation function was computed by averaging the correlation coefficients across subjects for each time lag. We performed bootstrapping with 1000 trials to establish the 95 % confidence interval (CI) for cross-correlations between the EEG signal and its corresponding speech envelope. We then separately calculated the 95 % CI for random correlations to estimate the chance level cross correlation for each of the spectral and TMR conditions, again using bootstrapping with 1000 trials. For this purpose, correlations were computed between randomly selected segments of the EEG signals and the speech envelope for each spectral condition at different time lags. Significant differences in latency or peak correlation (as described below) are defined as nonoverlapping 95 % CIs between two conditions, with a conservative assumption of $p < 0.05$ (Knol et al. 2011).

RESULTS

Comprehension Accuracy

A speech comprehension score was calculated as the percent of questions correctly answered for the attended story, for each spectral and TMR condition (Fig. 3). Two-way repeated measured ANOVA (5 spectral conditions $\times$ 2 TMRs) showed a significant spectral resolution main effect [$F_{4,28}=72.13$, $p < 0.001$] and a TMR main effect [$F_{1,7}=10.91$, $p = 0.013$], and no significant interaction between the two factors [$F_{4,28}=2.88$, $p = 0.08$]. Not surprisingly, there was a monotonic increase in comprehension scores as spectral resolution increased for both TMRs. Interestingly, speech comprehension was considerably poorer with 64-ch vocoder speech compared with unprocessed speech by 20 percentage points for both TMRs (3 dB: $t_7=6.01$, $p=0.001$; 0 dB: $t_7=4.6$, $p=0.002$).

Neural Responses to Attended and Unattended Speech

Ongoing EEG responses were recorded, while subjects attended to a narrated story presented in the presence of another narrated story read by a talker of the opposite gender. The EEG component tracking the speech envelope was extracted using DSS, and only the first DSS component was analyzed (Kong et al. 2014). For each spectral and TMR condition, correlation coefficients between the EEG response and the envelope of each talker were computed at different time lags. Figures 4 and 5 show the unproc and four vocoder conditions for attended and unattended speech at 3- and 0-dB TMR, respectively. For each function, the solid lines represent the correlation

Fig. 3. Average percent correct speech comprehension scores for different spectral and TMR conditions. Error bars represent one standard error around the mean.

coefficients at different time lags averaged across subjects. The shaded areas indicate 95 % CI from bootstrapping. The dashed lines represent the 95 % CI for random correlation. The latencies and peak correlations were determined as the local minima and local maxima for the first positive peak of the cross correlation ($P1^{XR}$) before 100 ms, the first negative peak ($N1^{XR}$) between 50 and 200 ms, and the second positive peak ($P2^{XR}$) between 100 and 400 ms.² To examine the differential cortical entrainments to the two speech streams, Figure 6 plots the *difference* in correlation coefficients between the attended and unattended speech (i.e., attended minus unattended) at different time lags for each TMR condition. The findings are as follows:

Effect of Spectral Resolution. The attentional modulation of neural responses was stronger for the higher than the lower spectral resolution conditions. For the higher spectral resolution conditions (unproc, 64 ch, and 32 ch), selective attention differentially modulated neural responses to the attended and unattended speech for both TMRs. In these conditions, correlations between attended and unattended speech were significantly different ($p < 0.05$; *nonoverlapping 95 % CI*) at both earlier (≤ 150 ms) and later (> 150 ms) latencies (cf. Figs. 4 and 5). Specifically, for the attended speech, we observed a prominent negative peak ($N1^{XR}$) around 100 ms and a positive peak ($P2^{XR}$) around 200 ms. For unattended speech, there was a positive peak ($P1^{XR}$)

around 50 ms and a negative peak ($N1^{XR}$) around 150–200 ms. The latency of the negative peak was consistently delayed for the unattended speech compared to the attended speech. The prominent positive peak around 200 ms for the attended speech, however, was not observed for the unattended speech for the vocoder conditions. These patterns of results are consistent with those recently reported by Kong et al. (2014) on a separate group (except for one subject) of normal-hearing listeners listening to naturally recorded speech at 6- and 0-dB TMRs.³

For the lower spectral conditions (16 and 8 ch), differences between the neural responses to attended and unattended speech were smaller compared to the higher resolution conditions. The correlation functions for the attended speech generally resembled the correlation functions of the unattended speech in the higher resolution conditions, particularly at 0-dB TMR, with a small positive peak around 50 ms. The prominent $N1^{XR}$ and $P2^{XR}$ for the attended speech in the higher spectral resolution conditions were not observed in severely degraded speech with 16 and 8 channels.

Tables 2 and 3 summarize the latency and absolute amplitude for the $N1^{XR}$ and $P2^{XR}$ responses for the attended speech and the $P1^{XR}$ and $N1^{XR}$ responses for unattended speech. It is noted that the peak latencies for some of the vocoder conditions (e.g., 8 and 16 ch) are not included in Table 2 because the peaks cannot be reliably identified in the

² As discussed by Horton et al. (2013), the positive and negative peaks in the cross-correlation functions are related to the $P1$ - $N1$ - $P2$ in the traditional EEG response to short discrete stimuli. Here, we used XR (stands for cross correlation) to indicate the cross correlation peaks, distinguishing from traditional EEG $P1$ - $N1$ - $P2$ components.

³ This indicates that the neural data is highly reproducible. Using different EEG recording equipment (G.tec in Kong et al. 2014; BrainVision in the current study) on different groups of subjects, the patterns of results for the unprocessed 0 dB TMR condition in the current study are very similar to those in the same test condition reported in Kong et al. (2014). These two sets of data are highly correlated for both the attended ($r_{301} = 0.9461$, $p < 0.001$) and unattended ($r_{301} = 0.9176$, $p < 0.001$) speech stream.

Fig. 4. Correlation coefficients between the EEG and the speech envelope for the attended (red) and unattended (blue) speech at different time lags in the five spectral conditions at 3-dB TMR. Shaded areas denote 95 % confidence interval (CI) of the cross-correlations. Dashed lines represent the 95 % CI for random correlation.

group data (i.e., peak correlation is not significantly different from random correlation). For the attended speech, we observed a trend that $N1^{XR}$ and $P2^{XR}$ latencies increased and $N1^{XR}$ and $P2^{XR}$ amplitudes decreased with decreasing spectral resolution. For example, for the higher spectral resolution conditions where $N1^{XR}$ and $P2^{XR}$ for the attended speech were clearly identified, their latencies were significantly longer ($p < 0.05$; *nonoverlapping 95 % CI*) for the 32- and 64-ch conditions compared to unprocessed speech for both TMR conditions, and the magnitudes of the correlations for $N1^{XR}$ and $P2^{XR}$ were also significantly smaller ($p < 0.05$; *nonoverlapping 95 % CI*)

for the 32-ch condition compared to unprocessed speech for at 0-dB TMR. The $N1^{XR}$ and $P2^{XR}$ had smaller amplitudes (not even detectable in some subjects) in the lower resolution than in the higher resolution conditions. For unattended speech, the $P1^{XR}$ latency was significantly ($p < 0.05$; *nonoverlapping 95 % CI*) increased, and the $N1^{XR}$ amplitude was significantly ($p < 0.05$; *nonoverlapping 95 % CI*) decreased, from the unproc condition compared to all other vocoder conditions. Taken together, these results suggest that degradation of spectral information affects the neural responses for both the attended and unattended speech.

Fig. 5. Correlation coefficients between the EEG and the speech envelope for the attended (red) and unattended (blue) speech at different time lags in the five spectral conditions at 0-dB TMR. Shaded areas denote 95 % confidence interval (CI) of the cross-correlations. Dashed lines represent the 95 % CI for random correlation.

Fig. 6. The difference in correlation between attended and unattended speech (attended minus unattended) across time lags for different spectral conditions at 3- (green) and 0-dB (purple) TMR. Shaded areas denote 95 % confidence interval (CI) of the difference in cross-correlations.

Relationship to Speech Comprehension. Figure 7 compares the NMIs and speech comprehension scores to characterize if speech comprehension is related to how strongly top-down attention modulates neural responses. The top panels in Fig. 7 compare the averaged speech comprehension score (left) and a NMI (right). The NMI follows the monotonic trend of comprehension scores with increasing spectral resolution and TMR strikingly well. A two-way repeated measured ANOVA (5 spectral conditions $\times$ 2 TMRs) was performed on the NMI data. Similar to the speech comprehension scores, the neural data showed a significant spectral condition main effect [$F_{4,28}=11.16$, $p=0.002$] and a TMR main effect [$F_{1,7}=15.51$, $p=0.006$], and no significant interaction between the two factors [$F_{4,28}=2.02$, $p=0.168$]. There was also a monotonic increase in NMI as spectral resolution increased for both TMRs, similar to the patterns observed in speech comprehension.

The bottom panels in Fig. 7 show the relationship between comprehension scores and NMIs. The panels include data for each test condition from individual subjects (small dots; 8 subjects $\times$ 5 spectral conditions $\times$ 2 TMRs = 80 data points), as well as the averaged data (larger symbols) across subjects for each spectral condition (left) and across spectral conditions for each subject (right). Normalization (i.e., z-score conversion) was performed prior to correlational analysis (Fischer and Milfont 2010). In the left panel, data from each subject was normalized to emphasize the difference between spectral conditions. In the right panel, data from each spectral condition was normalized to emphasize the difference between subjects. Combining the data from both TMRs, Pearson correlation confirmed that comprehension scores and NMIs were significantly correlated for

individual spectral conditions ($r_{80}=0.62$, $p<0.001$, left panel) as well as for individual subjects ($r_{80}=0.48$, $p<0.001$, right panel). This suggests a significant relationship between the magnitude of top-down attentional neural modulation and the degree of sound source segregation manipulated by differences in spectral resolution.

Effect of TMR. There was a trend that neural responses were different between 3- and 0-dB TMR across vocoder channel conditions. First, the difference in neural responses to attended and unattended speech was greater for the 3-dB than for the 0-dB TMR with decreasing spectral resolution (cf. Fig. 6). Second, the amplitudes of the peak correlations were generally greater for the higher TMR condition (Table 3). For vocoder speech, the mean $N1^{XR}$ amplitude tended to be larger for the 3-dB TMR than for the 0-dB TMR for both the attended and unattended speech. Also, the mean $P2^{XR}$ amplitude of the attended speech was smaller at 0-dB TMR than at 3-dB TMR.

DISCUSSION

In the present study, we systematically investigated the effect of spectral degradation on attentional modulation of neural responses to the attended and unattended speech streams. Subjects were tested with unprocessed speech and four vocoder channel conditions (i.e., 8, 16, 32, and 64 frequency bands). Our main findings are the following: (1) Differential attentional modulation of neural responses to the two speech streams diminished with decreasing spectral resolution, and (2) there was a significant relationship between differential attentional neural modulations and speech comprehension.

Table 2

The latency of the peaks in the cross-correlation functions for the attended and unattended speech in different listening conditions

<i>Attention</i>	<i>Spectral condition</i>	<i>TMR</i>	<i>Mean (95 % CI) (ms)</i>	<i>Mean (95 % CI) (ms)</i>
Attended speech	Unproc	3 dB	N1 ^{XR} 90 (82–96)	P2 ^{XR} 176 (170–184)
		0 dB	92 (84–98)	186 (180–190)
	64 ch	3 dB	114 (106–122)	210 (202–218)
		0 dB	124 (118–128)	218 (210–224)
	32 ch	3 dB	106 (96–120)	208 (194–216)
		0 dB	116 (106–135)	228 (208–260)
	16 ch	3 dB	–	–
		0 dB	–	–
	8 ch	3 dB	–	–
		0 dB	–	–
	Unproc	3 dB	P1 ^{XR} 48 (36–54)	N1 ^{XR} 130 (122–140)
		0 dB	46 (30–56)	134 (122–148)
Unattended speech	64 ch	3 dB	74 (58–82)	–
		0 dB	76 (54–88)	–
	32 ch	3 dB	80 (72–90)	–
		0 dB	76 (60–84)	–
	16 ch	3 dB	74 (66–82)	–
		0 dB	82 (76–94)	–
	8 ch	3 dB	64 (56–72)	–
		0 dB	70 (20–76)	–

Effect of Spectral Degradation on Attentional Modulation of Neural Responses

Consistent with previous electrophysiological findings (Ding and Simon 2012a; Horton et al. 2013; Kong et al. 2014; O’Sullivan et al. 2014), selective attention differentially modulates the neural responses to attended and unattended speech streams for highest resolution spectral conditions (i.e., unproc and 64-ch vocoder) in a challenging 0-dB TMR condition. We argue that high spectral resolution can support sound segregation and, as a result, selective attention can operate on the individually perceived auditory objects, enhancing the neural representation of the attended speech and suppressing the neural representation of unattended speech (Horton et al. 2013; Kong et al. 2014). The brain can do so by using temporal coherence to bind together acoustic features belonging to the same speech stream (Shamma et al. 2011) or by predicting which moments contain more information about the attended speech than the competing speech (Schroeder et al. 2008; Schroeder and Lakatos 2009; Zion-Golumbic et al. 2013).

However, for the lower-resolution spectral conditions (16-ch vocoder and 8-ch vocoder), particularly at the more challenging 0-dB TMR condition where the two speech streams have equal amplitude, no strong attentional modulation was observed. This suggests that the severely degraded speech does not support stream segregation, which in turn hinders the lis-

teners’ ability to selectively attend to the target stream and suppress the competitor. The neural response to both speech streams in these conditions is almost identical with only one peak at around 50 ms. This result is somewhat surprising because one may expect neural entrainment to the temporal envelope of the speech mixture with a positive peak at around 200 ms in the cross-correlation function, as observed for continuous speech presented in quiet in a passive listening condition (Kong et al. 2014). Interestingly, this response pattern instead resembles that for the unattended speech in the high-resolution conditions. This phenomenon may suggest that when listeners could not reliably follow the individual speech stream, they may either suppress both speech streams in the perceived mix or they somehow switch their attention back and forth between the target and masker at different points in time if they mistake the masker as the target due to poor voice pitch perception. Similarly, in a previous study by Ding et al. (2014), the MEG response to vocoded speech presented in quiet also became dominated by an early response near 50 ms when spectral resolution reduces.

We also observed an improvement in story comprehension and an increase in attentional modulation of neural responses when the spectral resolution increased from the 32- and 64-ch conditions to the unproc condition. This suggests that even 32- or 64-ch vocoder processing does not fully support sound segregation. This is consistent with previous work that showed less than optimal performance on speech

Table 3

The amplitude (i.e., the absolute maximum correlation value) of the peaks in the cross-correlation functions for the attended and unattended speech in different listening conditions

Attention	Spectral condition	TMR	Mean (95 % CI) (<i>r</i>)	Mean (95 % CI) (<i>r</i>)
Attended speech	Unproc	3 dB	N1 ^{XR} 0.07 (0.06–0.10)	P2 ^{XR} 0.07 (0.05–0.10)
		0 dB	0.06 (0.05–0.09)	0.07 (0.05–0.09)
	64 ch	3 dB	0.06 (0.04–0.08)	0.06 (0.04–0.08)
		0 dB	0.06 (0.04–0.07)	0.05 (0.04–0.06)
	32 ch	3 dB	0.04 (0.02–0.07)	0.05 (0.03–0.07)
		0 dB	0.03 (0.01–0.05)	0.03 (0.01–0.05)
	16 ch	3 dB	0.01 (0.00–0.03)	0.02 (0.01–0.03)
		0 dB	0.00 (0.00–0.02)	0.01 (0.01–0.03)
	8 ch	3 dB	0.01 (0.00–0.03)	0.02 (0.02–0.04)
		0 dB	0.01 (0.00–0.02)	0.00 (0.01–0.02) ^a
	Unattended speech	Unproc	P1 ^{XR} 0.05 (0.04–0.07)	N1 ^{XR} 0.05 (0.04–0.06)
			0.05 (0.03–0.06)	0.04 (0.03–0.05)
		64 ch	0.04 (0.03–0.05)	0.02 (0.01–0.04)
			0.03 (0.02–0.05)	0.01 (0.00–0.03)
		32 ch	0.06 (0.04–0.09)	0.03 (0.01–0.06)
			0.04 (0.02–0.05)	0.01 (0.01–0.03)
		16 ch	0.05 (0.03–0.07)	0.02 (0.01–0.03)
			0.04 (0.02–0.06)	0.00 (0.01–0.02)
		8 ch	0.04 (0.02–0.06)	0.01 (0.01–0.03)
			0.03 (0.01–0.05)	0.00 (0.00–0.02)

^a $p < 0.05$ using the most conservative estimates by Knol et al. (2011) regarding nonoverlapping 95 % CI between two TMR conditions

perception in a competing background and voice pitch perception with high-channel vocoder stimuli (e.g., Qin and Oxenham 2005; Kong and Zeng 2006; Massida et al. 2011). Qin and Oxenham (2005) showed that even though fundamental frequency (F0) difference limen was reasonably low (<1 semi-tone) with 24-ch vocoder processing, normal-hearing listeners were unable to benefit from F0 differences between the competing vowels in a concurrent-vowel paradigm, possibly attributing to the poor representation of the lower-order harmonics. Kong and Zeng (2006) showed that identification of Mandarin tones in noise was significantly poorer with 32-ch vocoder processing than with unprocessed speech, and Massida et al. (2011) demonstrated that voice discrimination was significantly reduced even with 64-channel vocoder processing compared to unprocessed speech.

Effect of TMR on Attentional Modulation of Neural Responses

Previous studies (Ding and Simon 2012a; Kong et al. 2014) found that, in a complex auditory scenario consisting of two competing speakers, cortical responses are largely invariant to the intensity ratio between the speakers when it varies in a range where the listeners can successfully attend to the target speaker. However, in the present study, we observed a TMR effect for the neural responses across vocoder conditions, which may imply that the listeners actively

use sound intensity as a cue in this experiment. For example, in the 8- and 16-ch conditions where sound segregation is severely impaired, the neural responses showed a weak P2^{XR} for the attended speech at 3-dB TMR but no clear P2^{XR} at 0-dB TMR. This may suggest that, with spectrally degraded signals that reduce listeners' ability to segregate individual speech streams, cortical representation limits to the glimpses of the attended speech at the moments in time and/or frequency only when the TMR is favorable. In this situation, neural responses phase lock to the peak of the envelope of the speech mixture, which is dominated by the envelope of the target speech when the TMR is positive.

Given that subjects were always tested with the 3-dB TMR condition first for each spectral condition, it is important to observe a degree of caution in interpreting these findings. It could be that the listeners habituated to the stimulation during the course of the test session. However, we argue that this explanation is unlikely because (1) we found no difference in patterns of neural responses between the averaged data for the first five repetitions and the second five repetitions of the same speech mixture, (2) we found no significant difference between TMR conditions for the unprocessed speech (also see Ding and Simon 2012a; Kong et al. 2014), but this was not the case for the vocoder speech with lower spectral resolution, (3) the stimuli were different between the 3- and 0-dB TMR conditions, and (4) a mandatory

Fig. 7. Comparison between behavioral speech comprehension performance and neural responses. *Upper panels* show comprehension scores (*left*, data replotted from Fig. 1) and NMI for different spectral conditions at 3- (red) and 0-dB (blue) TMR. Error bars represent one standard error around the mean. The *left lower panel* shows the relationship between comprehension score and neural index for different spectral conditions. In this analysis, to focus on the differences between spectral conditions, data from each subject (both behavioral and EEG) is normalized by using the z-score. Both individual subject data for each spectral condition (80 data points; *small dots*) and the averaged data for each spectral condition at each

TMR (*solid symbols*) are shown. Note that the *symbol shapes* are matched to the spectral conditions indicated in the *upper panels*. The *right lower panel* shows the relationship between comprehension score and NMI for individual subjects. In this analysis, to focus on the differences between subjects, data from each spectral condition (both behavioral and EEG) is normalized by using the z-score. Both individual subject data for each spectral condition (80 data points; *small black dots*) and the averaged data for each subject (*asterisks*) are shown.

break of at least 10 min was given to each subject after each block of testing.

Cortical Entrainment to the Speech Envelope

The current study utilized the envelope tracking of cortical response as a tool to study neural segregation of competing continuous speech streams. Although this response is usually referred to as the envelope tracking response by convenience, it does not merely reflect neural encoding of the stimulus envelope. That is, the strength of the cortical entrainment to the stimulus envelope is also influenced by other acoustic properties of the auditory stimuli. The current study and previous studies (Peelle et al. 2013; Ding et al. 2014) demonstrated that the so-called “envelope entrainment” strongly depends on the spectral and temporal fine structure. When the fine structure is intact and the auditory system is able to segregate concurrent speech streams, cortical activity is differentially entrained to the speech envelopes of different speakers, showing “object-level” neural encoding. Only when stream segregation

fails, cortical activity is entrained to the actual envelope of the speech mixture.

Relationship Between Comprehension and Neural Responses

A number of studies have examined the relationship between neural entrainment to the speech envelope and speech comprehension (see Ding and Simon 2014 for review). In auditory scenes consisting of competing speakers, previous work has studied cortical activities at high speech comprehension levels (>80 %) (e.g., Ding and Simon 2012a; Horton et al. 2013; Kong et al. 2014; O’Sullivan et al. 2014). O’Sullivan et al. (2014) reported a significant relationship between speech comprehension and EEG responses to the attended speech around 150–250-ms time lags. A systematic manipulation of spectral resolution in the present study produced a wide range of individual speech comprehension scores ranging from 0 % in 8-ch vocoder 0-dB TMR condition to 92 % in unprocessed 3-dB TMR condition. We observed a strong correlation between the NMI and the speech comprehension score. However, as discussed by Ding and Simon (2014), it is yet to be determined if

the changes in cortical entrainment arise from changes in speech perception or from changes in the acoustic properties of the degraded speech signals. In other words, it is unclear whether speech perception ability modulates neural responses or auditory encoding reflected by cortical entrainment influences downstream language processing. Furthermore, it is also observed that the EEG NMI explains part of the variance in the speech comprehension performance, indicating the possibility of using EEG to objectively assess speech perception performance in a competing background.

CONCLUSIONS

The present study investigated the effect of spectral degradation on differential neural tracking of the attended and unattended speech streams in a complex listening situation. Our results showed the detrimental effect of spectral degradation (via channel vocoding) on the cortical entrainment to different streams in the speech mixture. Specifically, attentional modulation of cortical entrainment was stronger for higher spectral resolution conditions than for lower spectral resolution conditions, indicating that degraded sensory inputs severely degrade neural stream segregation and auditory object formation, on which selective attention operates. When unable to follow individual streams with spectrally degraded speech, neural responses were affected by the intensity ratio between the attended and unattended speech. Furthermore, we showed a significant relationship between the attentional modulation of neural responses and speech comprehension. Taken together, these results illustrate cortical signatures of neural segregation of competing speakers for noise-vocoder speech and shed light on how cortical auditory representations are affected by reduced spectral resolution in complex listening environments. Importantly, our results demonstrate that severely spectrally degraded speech, as in cochlear-implant listening, does not support stream segregation, which negatively impacts listeners' ability to selectively attend to the target stream and suppress the competitor.

ACKNOWLEDGMENTS

We thank the reviewers and the associate editor for their helpful comments. This work is supported by NIH-NIDCD R01-DC012300 to Y.-Y.K.

Conflict of Interest

The authors declare that they have no conflict of interest.

REFERENCES

- BEST V, GALLUN FJ, CARLILE S, SHINN-CUNNINGHAM BG (2007) Binaural interference and auditory grouping. *J Acoust Soc Am* 121:420–432
- BREGMAN AS (1990) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge, MA
- BUSCHMAN TJ, MILLER EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315:1860–1862
- CULLING JF, DARWIN CJ (1993) The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Percept Psychophys* 54:303–309
- CULLING JF, SUMMERFIELD Q (1995) Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *J Acoust Soc Am* 98:785–797
- DARWIN CJ, CARLYON RP (1995) Auditory grouping. In: Moore BCJ (ed) *Hearing*. Academic Press, Orlando, FL, pp 387–424
- DARWIN CJ, HUKIN RW (2000A) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am* 107:970–977
- DARWIN CJ, HUKIN RW (2000B) Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *J Acoust Soc Am* 108:335–342
- DE CHEVEIGNE A, SIMON JZ (2008) Denoising based on spatial filtering. *J Neurosci Methods* 171:331–339
- DING N, SIMON JZ (2012A) Emergence of neural encoding auditory objects while listening to competing speakers. *Proc Natl Acad Sci U S A* 109:11854–11859
- DING N, SIMON JZ (2012B) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107:78–89
- DING N, SIMON JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8:311. doi:10.3389/fnhum.2014.00311
- DING N, CHATTERJEE M, SIMON JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88C:41–46
- ELHILALI M, FRITZ JB, CHI TS, SHAMMA SA (2007) Auditory cortical receptive fields: stable entities with plastic abilities. *J Neurosci* 27:10372–10382
- FISCHER R, MILFONT TL (2010) Standardization in psychological research. *Int J Psychol Res* 3:88–96
- FRITZ JB, ELHILALI M, DAVID SV, SHAMMA SA (2007) Auditory attention – focusing the searchlight on sound. *Curr Opin Neurobiol* 17:437–455
- GREENWOOD D (1990) A cochlear frequency-position function for several species – 29 years later. *J Acoust Soc Am* 87:2592–2605
- JASPER HH (1958) Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol* 10:370–375
- HORTON C, SRINIVASAN R, D'ZMURA M (2014) Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party.'. *J Neural Eng* 11:046015. doi:10.1088/1741-2560/11/4/046015
- HORTON C, D'ZMURA M, SRINIVASAN R (2013) Suppression of competing speech through entrainment of cortical oscillations. *J Neurophysiol* 109:3082–3093
- KERLIN JR, SHAHIN AJ, MILLER LM (2010) Attentional gain control of ongoing cortical speech representation in a "cocktail party.". *J Neurosci* 30:620–628
- KIDD G JR, ARBOGAST TL, MASON CR, GALLUN FJ (2005) The advantage of knowing where to listen. *J Acoust Soc Am* 118:3804–3815
- KONG Y-Y, ZENG F-G (2006) Temporal and spectral cues in mandarin tone recognition. *J Acoust Soc Am* 120:2830–2840

- KONG Y-Y, MULLANGI A, DING N (2014) Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hear Res* 316:73–81
- KONG Y-Y, DONALDSON G, SOMAROWTHU A (2015) Effects of contextual cues on speech recognition in simulated electric-acoustic stimulation. *J Acoust Soc Am* 137:2846–2857
- KNOL MJ, PESTMAN WR, GROBBEE DE (2011) The (mis)use of overlap of confidence intervals to assess effect modification. *Eur J Epidemiol* 26:253–254
- LALOR EC, POWER AJ, REILLY RB, FOXE JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J Neurophysiol* 102:349–359
- MASSIDA Z, BELIN P, JAMES C, ROUGER J, FRAYSSE B, BARONE P, DEGUINE O (2011) Voice discrimination in cochlear-implemented deaf subjects. *Hear Res* 275:120–129
- MESGARANI N, CHANG EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236
- MOORE BCJ (2007) Cochlear hearing loss: physiological, psychological and technical issues. John Wiley & Sons Ltd., West Sussex, UK
- OLDFIELD RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9:97–113
- O'SULLIVAN JA, POWER AJ, MESGARANI N, RAJARAM S, FOXE JJ, SHINN-CUNNINGHAM BG, SLANEY M, SHAMMA SA, LALOR EC (2014) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb Cortex*. doi:10.1093/cercor/bht355
- OXENHAM AJ (2008) Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. *Trends Amplif* 12:316–331
- PEELLE JE, GROSS J, DAVIS MH (2013) Phase-locking responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23:1378–1387
- POWER AJ, FOXE JJ, FORDE EJ, REILLY RB, LALOR EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur J Neurosci* 35:1497–1503
- QIN MK, OXENHAM AJ (2005) Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear Hear* 26:451–460
- RIMMELE JM, ZION GOLUMBIC E, SCHROEDER E, POEPEL D (2015) The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex* 68:144–154
- SCHROEDER CE, LAKATOS P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci* 32:9–18
- SCHROEDER CE, LAKATOS P, KAJIKAWA Y, PARTAN S, PUCE A (2008) Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* 12:106–113
- SHAMMA SA, ELHILALI M, MICHEYL C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34:114–123
- SHANNON RV, ZENG F-G, KAMATH V, WYGONSKI J, EKELID M (1995) Speech recognition with primarily temporal cues. *Science* 270:304–304
- SHINN-CUNNINGHAM BG (2008) OBJECT-BASED AUDITORY AND VISUAL ATTENTION. *TRENDS COGN SCI* 12:182–186
- SHINN-CUNNINGHAM BG, BEST V (2008) Selective attention in normal and impaired hearing. *Trends Amplif* 12:283–299
- STICKNEY GS, ZENG FG, LITOVSKY R, ASSMANN P (2004) Cochlear implant speech recognition with speech maskers. *J Acoust Soc Am* 116:1081–1091
- 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