Behavioral/Cognitive

Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech

Nai Ding[1] **and Jonathan Z. Simon**[1,2]

[1]Department of Electrical and Computer Engineering and [2]Department of Biology, University of Maryland, College Park, Maryland 20742

Speech recognition is remarkably robust to the listening background, even when the energy of background sounds strongly overlaps with that of speech. How the brain transforms the corrupted acoustic signal into a reliable neural representation suitable for speech recognition, however, remains elusive. Here, we hypothesize that this transformation is performed at the level of auditory cortex through adaptive neural encoding, and we test the hypothesis by recording, using MEG, the neural responses of human subjects listening to a narrated story. Spectrally matched stationary noise, which has maximal acoustic overlap with the speech, is mixed in at various intensity levels. Despite the severe acoustic interference caused by this noise, it is here demonstrated that low-frequency auditory cortical activity is reliably synchronized to the slow temporal modulations of speech, even when the noise is twice as strong as the speech. Such a reliable neural representation is maintained by intensity contrast gain control and by adaptive processing of temporal modulations at different time scales, corresponding to the neural δ and θ bands. Critically, the precision of this neural synchronization predicts how well a listener can recognize speech in noise, indicating that the precision of the auditory cortical representation limits the performance of speech recognition in noise. Together, these results suggest that, in a complex listening environment, auditory cortex can selectively encode a speech stream in a background insensitive manner, and this stable neural representation of speech provides a plausible basis for background-invariant recognition of speech.

Introduction

Speech recognition is robust with respect to various listening backgrounds. The slow temporal modulations (<16 Hz) that constitute the speech envelope (Rosen, 1992) contribute to robust speech recognition in two important ways. First, they reflect the syllabic and phrasal rhythm of speech (Greenberg et al., 2003) and, in quiet listening environments, lead to high intelligibility with even very coarse spectral information (Shannon et al., 1995). Accordingly, it has been proposed that cortical activity synchronized to the speech envelope underlies the parsing of speech into basic processing units (e.g., syllables) (Giraud and Poeppel, 2012). Second, in complex auditory scenes, slow temporal modulations provide cues to group features belonging to the same sound stream (Shamma et al., 2011), and therefore selective neural synchronization to a speech stream has been hypothesized as a mechanism to segregate the speech stream from the listening background (Schroeder and Lakatos, 2009; Shamma et al., 2011). Both the segregation of speech from background and the parsing of speech into perceptual units are prerequisites for robust speech recognition. Therefore, if cortical synchronization to the speech envelope is causally involved in these processes, it must reliably occur in any listening environment that does not extinguish speech intelligibility. This critical prediction is tested in this study.

An acoustic background interferes with speech in two ways: via energetic masking and informational masking (Brungart, 2001; Scott et al., 2004). Recently, it has been shown that cortical synchronization to speech is robust to strong informational masking caused by an interfering speech stream (Kerlin et al., 2010; Ding and Simon, 2012b; Mesgarani and Chang, 2012). Here, we further test whether it is also robust to energetic masking caused by spectrotemporal overlap between the energy of speech and any acoustic background. Strong energetic masking caused by (e.g., stationary noise) can produce severe degradation in speech encoding at the level of the auditory nerve (Delgutte, 1980) and brainstem (Anderson et al., 2010), but how these degraded neural representations are rescued by the higher level auditory system is not well understood.

The current study investigates the cortical encoding of speech embedded in spectrally matched stationary noise, the most classic example for energetic masking (Festen and Plomp, 1990). The neural recordings were obtained using MEG from subjects listening to a spoken narrative mixed with noise at different signal-to-noise ratios (SNRs). Spectrally matched stationary noise reduces the intensity contrast of the speech and distorts the spectrotemporal modulations (Fig. 1*A*,*B*). Under such strong acoustic interference, psychoacoustic studies suggest that robust speech recognition arises from listeners' insensitivity to stimulus intensity contrast (Stone et al., 2011) and selective processing of the temporal modulations with rates less corrupted by noise

Received Nov. 14, 2012; revised Feb. 6, 2013; accepted Feb. 17, 2013.

Author contributions: N.D. and J.Z.S. designed research; N.D. and J.Z.S. performed research; N.D. and J.Z.S. contributed unpublished reagents/analytic tools; N.D. and J.Z.S. analyzed data; N.D. and J.Z.S. wrote the paper.

This work was supported by the National Institute of Deafness and Other Communication Disorders Grant R01-DC-008342.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Jonathan Z. Simon, Department of Electrical & Computer Engineering, University of Maryland, College Park, MD, 20742. E-mail: jzsimon@umd.edu.

DOI:10.1523/JNEUROSCI.5297-12.2013

0270-6474/13/335728-08$15.00/0

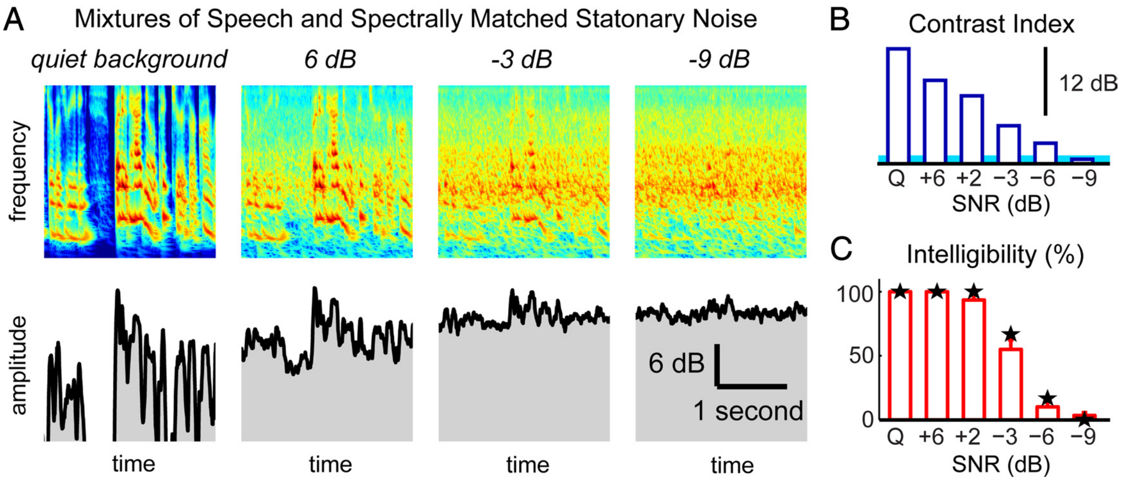

Figure 1. Speech embedded in spectrally matched stationary noise. ***A***, The auditory spectrogram (top) and the broadband temporal envelope (bottom) of speech embedded in noise, at 4 SNRs. The background noise causes severely degradation to the spectro-temporal features of speech (in this illustration but not in the experiment, the same speech segment is used in every SNR condition). ***B***, The contrast index characterizes the spectro-temporal contrast of the stimulus at each SNR. The shaded blue area covers the fifth to 95th percentile of the contrast index calculated for stationary noise alone, and the SNR condition Q indicates a quiet background. The intensity contrast of the stimulus decreases continuously with SNR. ***C***, Subjectively rated intelligibility of speech (bars), and percentage of comprehension questions correctly answered (★). The intelligibility remains unaffected by SNR until −3 dB SNR.

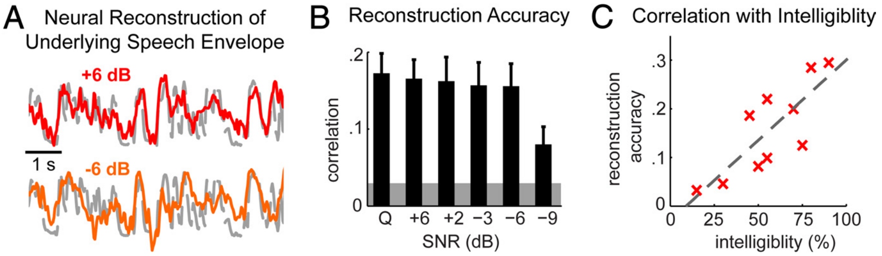

Figure 2. Neural reconstruction of the temporal envelope of speech. ***A***, The red and orange waveforms are the envelopes reconstructed from the neural responses in two sample SNR conditions. The dashed gray waveform is the envelope of the underlying speech in each stimulus. The neural construction matches the speech envelope well at both SNRs. The neural reconstructions illustrated are averaged over trials and subjects (*N* = 5, for the increasing and decreasing SNR conditions, respectively). ***B***, Correlation between the single-trial neural reconstruction and the envelope of the underlying speech, as a function of SNR. The correlation is averaged over trials. Error bar indicates SEM over subjects. The 95th percentile of chance level reconstruction accuracy is shaded (permutation test). ***C***, Relationship between the neural reconstruction accuracy and speech intelligibility, at −3 dB SNR. Each subject is shown by a red cross. The neural and behavioral results are highly correlated, with the regression line shown by the dashed line.

(Jorgensen and Dau, 2011). Here, we test whether these computational strategies are indeed implemented in the human brain, via, for example, neural adaptation to the mean and variance of stimulus intensity (Robinson and McAlpine, 2009) and stimulus-dependent neural encoding of temporal modulations (Escabí et al., 2003; Woolley et al., 2006; Lesica and Grothe, 2008).

Materials and Methods

Subjects. Eleven right-handed (Oldfield, 1971) young adults (7 females) between 20 and 31 years old participated in the experiment. All reported normal hearing. One subject was excluded because of the lack of auditory responses to both tones and speech. Subjects were paid for their participation. The experimental procedures were approved by the University of Maryland institutional review board. Written informed consent was obtained from each subject before the experiment.

Stimuli and procedure. The stimuli were taken from the beginning of a narration of the story *Alice's Adventures in Wonderland* (http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-4/). The sound recording was low-pass filtered <4 kHz and divided into twelve 50-s duration sections, after long speaker pauses (> 300 ms) were shortened to 300 ms. A spectrally matched stationary noise was generated based on a 12-order linear predictive model estimated from the speech recording and mixed into speech with one of six SNRs, that is, quiet (no noise added in), +6 dB, +2 dB, −3 dB, −6 dB, and −9 dB. The intensity of speech, measured by root mean square, was the same for all sections and the intensity of noise was varied to create different SNRs.

All the sections were presented sequentially and then repeated twice (3 trials total). The noise was frozen over trials (i.e., the same speech and noise mixture was used for every trial within a condition). Although each instance of frozen noise contained its own distinctive spectrotemporal features, any effects of those features were diluted over the 50 s duration of the stimulus. The subjects were asked a comprehension question after each section and also rated intelligibility of speech (in percentage) during the first presentation of each section. All stimuli were presented identically to both ears, and the subjects were required to close their eyes while listening.

The SNR either only decreased or only increased every two sections. For the decreasing SNR order (*N* = 5), no noise was added to the first two sections; noise 6 dB weaker than speech was added to the following two sections, and then the noise level kept increasing over the remaining sections. The increasing SNR order (*N* = 5), in contrast, started with the lowest SNR (i.e., −9 dB) and finished with the quiet condition. The story continued naturally under either presentation order. As a result, the speech material used to create the speech-noise mixture was the same for the +2 dB SNR condition in the decreasing SNR order (the second condition) and the −6 dB SNR condition in the increasing SNR order (also the second condition). In Figure 2*A*, because the waveforms of the neural reconstruction were shown, the subjects were grouped separately based on stimulus presentation order. The SNR order affects neither speech intelligibility (SNR × order, two-way repeated-measures ANOVA) nor the neural reconstruction of speech (SNR × order × trial, three-way repeated-measures ANOVA), and therefore was not distinguished in any analyses and figures other than Figure 2*A*. In summary, each SNR condition consisted of two 50 s duration sections, each repeated 3 times. In the MEG analysis, responses from the two sections with the same SNR were concatenated.

Before the main experiment, 100 repetitions of a 500 Hz tone pip were presented to elicit the M100 response, which is a reliable auditory response measured 100 ms after the onset of a tone pip and whose neural source is easy to localize within auditory cortex (Lütkenhöner and Steinsträter, 1998). The neuromagnetic signal was recorded using a 157-channel whole-head MEG system (KIT), with 1 kHz sampling rate. A 200 Hz low-pass filter and a notch filter at 60 Hz were applied online, and environmental noise was removed offline. The neural recordings were filtered offline between 1 and 9 Hz and down-sampled to 40 Hz. More details of the recording procedure are as described previously (Ding and Simon, 2012a).

Stimulus characterization. The auditory spectrogram of the stimulus was calculated using a subcortical auditory model (Yang et al., 1992) and expressed in linear amplitude scale. In the frequency by time auditory spectrogram, the frequency channels were logarithmically spaced and each time bin in the spectrogram was 5 ms in duration. In each frequency channel, the energy fluctuation over time is referred to as the narrowband envelope. The broadband envelope of stimulus was defined as the sum of the auditory spectrogram over frequency. The spectrotemporal contrast of a stimulus was characterized using a contrast index, the coefficient of variation of the auditory spectrogram, an extension of the fluctuation index (Nelken et al., 1999). The coefficient of variation is the SD of the amplitude of the auditory spectrogram over the 50 s stimulus divided by the mean. It is zero for a sound with its energy constant over time and frequency and grows as the contrast (i.e., depth) of the spectrotemporal modulations increases.

Neural reconstruction of stimulus. The temporal envelope of the speech-noise mixture, or the speech only, was reconstructed by linearly integrating MEG activity over time and sensors. The reconstructed speech envelope is expressed as follows:

$$\hat{E}(t) = \sum\nolimits_k \sum\nolimits_{0 < \tau \leq \ 500\text{ ms}} M_k(t + \tau) D_k(\tau),$$

where $M_k(t)$ is the MEG signal from a sensor k and $D_k(t)$ is the linear decoder for the same sensor. The envelope to reconstruct, $E(t)$, is either the envelope of the actual stimulus (the speech-noise mixture) or the envelope of the underlying speech (embedded in the stimulus). The decoder $D_k(t)$ was optimized using boosting with 10-fold cross-validation (David et al., 2007) to maximize the correlation between $\hat{E}(t)$ and $E(t)$. To reduce computational complexity, the MEG sensors in each hemisphere were compressed into 3 components using denoising source separation (de Cheveigné and Simon, 2008). Both hemispheres were used unless otherwise specified.

Intertrial correlation analysis. The phase locking of the neural response was investigated in narrow frequency bands (2 Hz wide) by calculating the intertrial correlation of the neural response. The intertrial correlation measures the reliability of neural responses when the same speech noise mixture is repeated, and is a reflection of the strength of phase-locked neural activity. Unlike the neural reconstruction analysis, which examines how accurately the speech envelope is encoded, the intertrial correlation analysis is sensitive to any phase-locked response elicited by the speech-noise mixture. The analysis is made possible by the use of frozen noise from trial to trial. The major component of MEG response was extracted using the first denoising source separation component (de Cheveigné and Simon, 2008) and applied to this analysis. The phase-locking spectrum of the neural response to speech has a low-pass shape (i.e., the precision of neural phase locking decreases over frequency) (Ding and Simon, 2012a). To estimate the low-pass cutoff frequency, the phase-locking spectrum is modeled using a sigmoidal function as follows:

$$1 - 1/\exp(-\alpha(f - f_T)).$$

The slope parameter α and location parameter f_T are fitted in the least-squares sense. In this model, because a sigmoidal function is bounded between 0 and 1, at each SNR the maximal intertrial correlation is normalized to 1 and the minimum is normalized to 0.

Temporal response function. The temporal response function (TRF) was obtained by deconvolving the continuous neural response evoked by the continuous speech stream, giving a waveform that can be interpreted as the response resulting from a unit power increase of the stimulus (Ding and Simon, 2012a). A TRF was estimated based on each MEG sensor, and the MEG data were averaged over trials in the TRF analysis. To estimate the TRF, a spectro-TRF is first estimated using boosting with 10-fold cross validation (David et al., 2007), using the procedure described by Ding and Simon (2012b). The TRF is obtained by summing the spectro-TRF over frequency. The M50_{TRF} was determined as the response peak between 0 and 80 ms, which has a magnetic field topography negatively correlated with that of the M100. The M100_{TRF} was determined as the response peak between 80 and 180 ms, which has a magnetic field positively correlated with that of the M100 (Ding and Simon, 2012b).

Neural source analysis. The neural sources of the M50_{TRF}, M100_{TRF}, and M100 (evoked by a tone pip) were modeled by an equivalent-current dipole (ECD) in each hemisphere, based on a spherical head model (Ding and Simon, 2012b). The median correlation between the fitted ECD magnetic field and the measured magnetic field is >90% in both hemispheres and for all the M50_{TRF}, M100_{TRF}, and M100. Comparing the ECD positions of different peaks in TRF, we included only ECDs successfully capturing the measured magnetic field, characterized by a >80% correlation between the ECD magnetic field and the measured magnetic field. Only one subject was excluded this way. After the ECD positions were determined, the moment of the dipole was estimated using the least-squares method (Mosher et al., 2003). For the dipole moment, the polarity of the M100_{TRF} is defined as negative, to be consistent with the polarity of the N1 peak of EEG. The TRF linearly projected to the ECD location was used to analyze the amplitude and latency of the M50_{TRF} and M100_{TRF} (Ding and Simon, 2012b).

Time-dependent speech reconstruction. In the stimulus reconstruction analysis, the decoder $D_k(t)$ integrates MEG activity over a 500 ms time period. The length of the period, however, can also be varied to investigate which time intervals carry more information. During this varying integration window analysis, however, the autocorrelation of the speech envelope must be taken into consideration. For example, the response at time $t = 50$ ms, $M(t - 50)$, should contain no information of the stimulus at a future time t, $E(t)$. Nevertheless, if $M(t - 50)$ encodes information of the stimulus at time $t = 100$ ms, which is correlated with $E(t)$, then from $M(t - 50)$ some information about $E(t)$ can be reconstructed, implicitly through $E(t - 100)$. Therefore, in the integration window analysis, we partialled out the autocorrelation of the envelope using an extended model as follows:

$$E(t) = \sum\nolimits_k \sum\nolimits_{1 \leq \tau \leq T} M_k(t + \tau) D_k(\tau) + \sum\nolimits_{1 \leq \tau \leq T} E(t - \tau) D_A(\tau) + \varepsilon(t),$$

where $\varepsilon(t)$ is the unexplained residual. $D_k(t)$ and $D_A(t)$, the decoder and the regressor for speech autocorrelation, are estimated simultaneously using boosting (David et al., 2007). The length of the time period integrated by the decoder, T, varies from 50 to 1000 ms, whereas the maximal time range where the autocorrelation of speech is considered, T^*, is set to 500 ms. In this case, the reconstructed neural response, $\hat{E}^*(t) = \Sigma_k \Sigma_{1 \leq \tau \leq T} M_k(t + \tau) D_k(\tau)$ is a reconstruction of the component in speech envelope that cannot be predicted by its own history because of the rhythm of speech (i.e., the unpredictable information at a given moment).

Results

Noise robust cortical reconstruction of speech

The stimulus consists of a narrated story that is divided into 50 s duration sections. Each is presented either in quiet (alone) or with added spectrally matched stationary noise (6 SNR levels ranging from −9 to +6 dB). A contrast index is used to characterize how the background noise reduces the intensity contrast (i.e., the depth of the spectrotemporal modulations) of the stimulus. As shown in Figure 1*B*, the intensity contrast of the speech-noise mixture decreases monotonically with decreasing SNR, until finally reaching the intensity contrast of stationary noise alone, at −9 dB SNR. The intelligibility of the stimulus starts to decrease at +2 dB and essentially disappear at −9 dB.

To investigate how the cortical representation of speech is affected by noise, we attempted to reconstruct the temporal en-

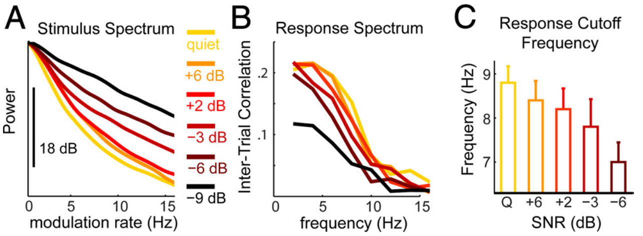

Figure 3. Neural encoding of temporal modulations. ***A***, The power spectrum of the stimulus envelope, at different SNRs. Each spectrum is normalized based on its power density at 0.1 Hz, to emphasize changes in shape rather than scale. The modulation spectrum of speech in quiet background (yellow) has the sharpest low-pass shape, and background noise increases the proportion of the stimulus power in higher modulation rates. ***B***, The phase-locking spectrum of the neural response, which quantifies the reliability of the neural response at each frequency under identical stimulus presentations using intertrial response correlation. The spectrum is consistently low-pass in shape but with a cutoff frequency that decreases with poorer SNR. ***C***, The cutoff frequency of the phase-locking spectrum (not reliably estimable at −9 dB SNR) decreases with SNR. Error bars indicate SEM over subjects.

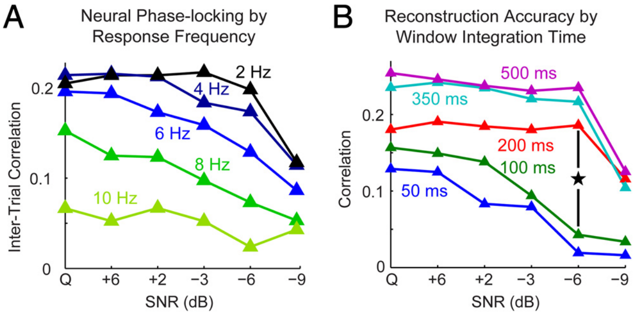

Figure 4. Stability of cortical synchronization to speech depends on long-term temporal integration. ***A***, The phase locking of neural activity as a function of SNR. When SNR decreases from +6 to −6 dB, the neural phase locking at very low frequencies (e.g., 2 Hz) is stable but the neural phase locking at higher frequencies (e.g., 8 Hz) immediately and continuously decreases, with intermediate trends of decrease at intermediate frequencies. ***B***, Correlation between the actual speech envelope and the envelope reconstructed based on the MEG response using temporal integration windows of different sizes. The reconstruction accuracy is stable >−6 dB SNR only when the temporal integration window is >200 ms. The largest change in reconstruction accuracy (★), which occurs at −6 dB SNR when the window size increases from 100 to 200 ms.

velope of the underlying speech (as opposed to the actual stimulus including noise), from the cortical response to the noisy stimuli (Fig. 2*A*). The accuracy of the reconstruction reflects how precisely cortical activity is synchronized to the speech envelope, even in the presence of background noise. It remains unaffected by the background noise until the noise is 9 dB stronger than speech, whereupon it drops. Above −9 dB, the stable neural reconstruction results do not follow the variable contrast index (Fig. 1*B*), and suggest a stable neural representation of speech maintained by contrast gain control.

At the intermediate SNR of −3 dB, the subjectively rated speech score varies broadly over subjects, with a median of 55%. At this SNR, individual speech scores are strongly correlated with the accuracy of neural reconstruction (Fig. 2*C*). The correlation coefficient is 0.79 ± 0.15 (mean ± SEM; the SEM is consistently used in the paper to describe subject variations and is calculated using bootstrap), significantly positive ($p < 0.005$, bootstrap). When the two hemispheres are analyzed separately, the reconstruction in each hemisphere is also correlated with speech intelligibility (mean correlation coefficient: 0.81, no significant difference between hemispheres, $p = 0.41$, bootstrap). At higher and lower SNR conditions, the speech scores clump near ceiling (median, >90%) or floor (≤10%) values, respectively (Fig. 1*C*), precluding analogous computations there.

To investigate whether the successful neural reconstruction of the underlying clean speech is a result of the neural encoding of the actual stimulus, we also reconstructed the envelope of the actual noisy stimulus from cortical activity. This reconstruction, although naively more straightforward, is less accurate than the reconstruction of the underlying speech for SNRs between +6 dB and −3 dB ($p < 0.01$, paired *t* test). Therefore, we see that auditory cortex predominantly synchronizes to the underlying speech rather than the physically presented sound mixture. The mechanisms underlying this robust neural representation are analyzed in the following sections.

Modulation sensitivity

Speech and noise each have a distinct modulation spectrum (the power spectrum of the temporal envelope), with the noise possessing more energy at higher modulation rates. Therefore, when noise is introduced, the energy of the stimulus envelope spreads into higher modulation rates (Fig. 3*A*). Consequently, if cortical activity were simply following the temporal modulations of the stimulus, it would also spread into higher frequencies. This conjecture, however, can be ruled out (Fig. 3*B*). Indeed, at the higher frequencies (e.g., near 7 Hz), the most reliable phase-locked response, measured by intertrial response correlation, is seen with a quiet acoustic background, and the response spectrum progressively shifts toward lower frequencies as more noise is introduced. It is worth emphasizing that Figure 3*B* shows the phase-locking of the measured neural response rather than the reconstructed speech, and therefore is sensitive to both the phase-locked response to speech and the phase-locked response to the frozen background noise.

The cutoff frequency of response spectrum (Fig. 3*C*, estimated by fitting each spectrum to a sigmoidal function) decreases monotonically from 8.7 ± 0.4 Hz to 7.0 ± 0.5 Hz as the SNR decreases from infinity (quiet background) to −6 dB. Between +6 dB and −6 dB, the cutoff frequency decreases 0.72 ± 0.29 Hz every 6 dB (linear regression). Therefore, as the noise level rises, the auditory system reduces its sensitivity to fast temporal modulations, allowing it disregard the increasingly stronger fast modulations introduced by the noise.

Temporal integration

An alternative measure of how the neural phase-locking depends on SNR and frequency is to analyze the response phase locking as a function SNR, at each frequency (Fig. 4*A*). At very low frequencies (e.g., 2 Hz), the response is not affected by noise until the lowest SNR of −9 dB. At higher frequencies (e.g., 6 and 8 Hz), however, the response degrades immediately and continuously with SNR. The SNR range within which intertrial correlation remains stable is further probed using a one-way ANOVA test. At 2 Hz, the intertrial correlation is not significantly affected by SNR from the quiet condition to −6 dB SNR ($p > 0.5$) but is affected by SNR if the −9 dB SNR is included ($p < 0.01$). Similarly, at 4 and 6 Hz, respectively, the intertrial correlation is stable until +2 and +6 dB ($p > 0.5$), but not any lower SNRs ($p < 0.01$). The stability of neural phase locking at lower, but not higher, frequencies suggests that the long-term temporal integration is important in maintaining a noise-robust neural representation.

To confirm the role of long-term integration in encoding speech envelope, we again applied the neural reconstruction analysis, but with a varying length time integration window. In the analysis shown in Figure 2, the reconstruction of the stimulus

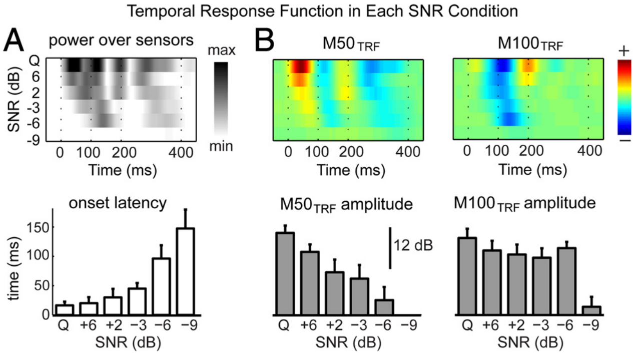

Figure 5. SNR-dependent temporal response function. ***A***, The instantaneous TRF power, summed over sensors. The TRFs from all SNR conditions are stacked vertically. The latency at which the TRF amplitude surpasses the noise floor is shown in the bottom. The TRF onset is significantly delayed by noise. ***B***, The TRFs at the neural sources of the $M50_{TRF}$ and $M100_{TRF}$ (top). The amplitude of the $M50_{TRF}$ decreases when the level of noise increases (compare with the stimulus contrast index illustrated in Fig. 1*B*), whereas the amplitude of the $M100_{TRF}$ remains stable until −9 dB SNR. Error bars indicate SEM over subjects.

at each time moment is based on the response in a 500 ms time window starting from that moment. When this window size is allowed to vary, the reconstruction results show a strong dependency on the integration time (Fig. 4*B*). At the poorer SNRs (e.g., −3 to −9 dB), the decoding results improve substantially when the window of integration is allowed to increase in size from 100 to 200 ms. In the −6 dB to 6 dB SNR range, the reconstruction accuracy is affected by SNR if the integration window is <200 ms ($p < 0.003$, one-way ANOVA) but not affected by SNR if the window is ≥200 ms ($p > 0.07$, one-way ANOVA). This demonstrates the importance of long-term (> 100 ms) integration in decoding speech in a strong noise background.

TRF

To explicitly characterize how the spectrotemporal features of the stimulus are encoded cortically as a function of time, and by cortical area, for each MEG sensor we estimate a TRF, which characterizes the time course of neural activity evoked by a unit power increase of the stimulus (Ding and Simon, 2012b). Although the neural reconstruction integrates responses over a specified duration, the TRF describes the neural response at each time lag between the stimulus and the response through deconvolution. In the TRF analysis, the intensity contrast of the stimulus is normalized separately for each SNR condition, to focus on stimulus-dependent response properties separate from the (dramatic) contrast gain control. With the stimulus thus normalized, an SNR-independent TRF amplitude would demonstrate a neural representation independent of the mean and variance (i.e., contrast) of the stimulus intensity.

The instantaneous TRF power, averaged over all MEG sensors, is shown in Figure 5*A*, top. The onset latency of the TRF (the earliest time point when the TRF amplitude passes the 99th percentile of the prestimulus TRF amplitude) is prolonged as the noise level rises (Fig. 5*A*, bottom). This latency elongation is statistically significant because the relationship between onset latency and SNR, when fitted by a line, has a significantly negative slope ($p < 0.001$, bootstrap). The earliest two components of the TRF, called the $M50_{TRF}$ and $M100_{TRF}$, are extracted and further analyzed.

A bilateral equivalent current dipole (ECD) based neural source localization shows that the ECD source location of the $M100_{TRF}$ is consistent with the ECD source location of the M100 evoked by a tone pip (no significant difference, $p > 0.3$, paired *t* test), whereas the ECD position of the $M50_{TRF}$ is on average 11 mm more anterior than that of the M100 in both hemispheres ($p < 0.02$ for the right hemisphere, $p < 0.003$ for the left hemisphere, paired *t* test). The ECD position of the $M50_{TRF}$ is also on average 10 mm (13 mm) more anterior than that of the $M100_{TRF}$ in the left (right) hemisphere (statistically significant in the right hemisphere only, $p < 0.02$, paired *t* test). The TRFs at the ECD position of $M50_{TRF}$ and $M100_{TRF}$ are shown (stacked vertically by SNR condition) in Figure 5*B*. The TRFs are averaged over the two hemispheres because very similar results are seen in each. The amplitude of the $M50_{TRF}$ decreases continuously with SNR, whereas the amplitude of the $M100_{TRF}$ is insensitive to SNR until the SNR decreases to −9 dB. A linear regression analysis shows that, in between −6 dB and 6 dB SNR, the amplitude of the $M50_{TRF}$ decreases 1.0 ± 0.2 dB (significantly negative, $p < 0.001$, bootstrap), whereas the amplitude of the $M100_{TRF}$ changes 0.0 ± 0.2 dB (not significantly) each 1 dB SNR change. The same regression analysis reveals that the latency of the $M50_{TRF}$ increases with decreasing SNR, with a change of 3.0 ± 0.6 ms/dB.

Temporal modulations within frequency channels

In auditory cortex, temporal modulations in different carrier frequency channels, called the narrowband envelopes, are represented by different neural populations, at least in tonotopically organized areas. These different populations, however, cannot be resolved using the current neural recording technique, and their responses are mixed in the MEG recording as a large-scale response following the broadband envelope of speech (Ding and Simon, 2012a). In the analyses above, the noise-robust neural representation contrasts the noise-sensitive broadband envelope. There is a possibility, however, that only the broadband envelope is vulnerable to noise, whereas the narrowband envelopes, which cortical neurons actually encode, are more robust. If this were the case, the noise-robust neural representation would naturally arise from the extraction of narrowband envelopes that occur in the cochlea rather than central mechanisms, such as contrast gain control and changes in modulation sensitivity. To rule out this possibility and to draw a possible link between the large-scale neural representation of the broadband envelope and the local neural network level representation of narrowband envelopes, in the following, we examine how the narrowband envelopes of speech are degraded by noise and whether the degradation is similar to the degradation to the broadband envelope.

As is shown by Figure 6*A*, as the level of the background noise rises, the narrowband envelopes in all carrier frequency channels are weakened, and the loss in power is more severe at lower modulation rates. This effect is consistent with what is observed for the broadband envelope although quantitatively weaker. Therefore, the fact that the neural response is not weakened by noise at low frequencies (Fig. 3*B*) cannot be the result of only selective encoding of the narrowband envelopes in some carrier frequency channels but requires contrast gain control within frequency channels.

The background noise reduces the dynamic range of the narrowband envelopes of speech and also distorts its shape. The shape distortion is quantified using the correlation between the envelope of a speech-noise mixture and the envelope of the original clean speech. As is shown in Figure 6*B*, the noise-induced distortion in the narrowband envelope is more severe than the distortion in the broadband envelope, for the stationary noise used in the current study. Furthermore, the noise-induced distortion is more severe at higher modulation rates, for both the

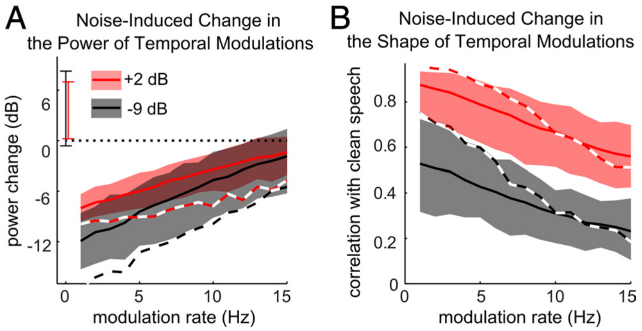

Figure 6. Noise-induced changes for the broadband and narrowband envelopes. ***A***, Noise-induced changes in the power spectrum of the envelopes. ***B***, Noise-induced changes in the shape of the temporal modulations. The changes in narrowband envelopes are shown by the solid lines as an average over carrier frequency channels between 160 Hz and 3600 Hz. The shaded area covers the fifth and the 95th percentile of the change in individual carrier frequency channels. The noise-induced change in the broadband envelope is shown by the dashed lines. The narrowband and the broadband envelopes are affected by the background noise in similar ways, showing both noise-induced reduction in modulation power and distortions in shape.

broadband and narrowband envelopes. Therefore, within carrier frequency channels, reducing sensitivity to faster modulations is still a valid strategy to maintain a robust neural representation of speech. In summary, spectrally matched stationary noise affects the narrowband envelope and the broadband envelope of speech in qualitatively similar ways. Consequently, neural computations suggested in the current study, such as contrast gain control and adaptive processing of temporal modulations, are likely to reflect computational processes occurring in local neural networks, rather than mechanisms, such as selective encoding of certain carrier frequency channels.

Discussion

By recording from human subjects listening to continuous speech corrupted by a noise background, this study demonstrates a background-insensitive neural representation of speech in auditory cortex. The neural entrainment to very slow temporal modulations (<4 Hz) of speech remains stable until the noise background is more than twice as strong as the speech (−9 dB SNR). The neural entrainment to relatively faster temporal modulations (4–8 Hz), however, appears to be more sensitive to noise. Mechanistically, two distinct types of acoustic degradation caused by noise background (i.e., the compression of stimulus dynamic range and the distortion of fast temporal modulations) are separately compensated for in the auditory system by contrast gain control and a shift in modulation sensitivity.

Background invariant neural representations of speech

Here, we demonstrate that the neural synchronization to speech in human auditory cortex is resilient to the strong energetic masking of stationary noise. Previous studies have shown that the neural synchronization is also resilient to the strong informational masking of a competing speech stream (Ding and Simon, 2012b; Mesgarani and Chang, 2012). Considering these results together, it is now demonstrated that the cortical encoding of the slow modulations of speech is robust to both energetic and informational masking, the two broad categories of acoustic interference to speech, and therefore likely to occur in any listening condition that allows speech recognition. This indicates that auditory scene analysis (e.g., the segregation of an auditory object from the acoustic background) is generally achieved at the level of auditory cortex (Griffiths and Warren, 2004; Bar-Yosef and Nelken, 2007; Shinn-Cunningham, 2008; Fishman and Steinschneider, 2010; Shamma et al., 2011), and the auditory object of the listener's interest is selectively represented.

This robust neural encoding of slow temporal modulations is only achievable by complex neural computations, including what can be characterized as contrast gain control and long-term temporal integration, as will be discussed in the following.

Contrast gain control in auditory cortex

The dynamic range of speech is severely compressed by acoustic degradation caused by background noise and, for example, reverberation. The loss of dynamic range, however, has little effect on speech recognition, especially if the shape of the temporal modulations is maintained (Stone et al., 2011), indicating an auditory representation insensitive to dynamic range. Indeed, in single-neuron studies with nonspeech stimuli, neural adaptation to the mean and/or variance of sound intensity has been observed and stronger gain control effects are seen along the ascending auditory pathway (Dean et al., 2005; Robinson and McAlpine, 2009; Watkins and Barbour, 2009; Wen et al., 2009; Zilany et al., 2009; Rabinowitz et al., 2011).

In this study, a hierarchy of contrast gain control is seen in auditory cortex. The early $\mathrm{M50_{TRF}}$ component is significantly weakened as the dynamic range of the stimulus is compressed by background noise, reflecting incomplete contrast gain control. The neural source location of the $\mathrm{M50_{TRF}}$ is approximately consistent with core auditory cortex because it is more anterior to the neural sources of the M100 and $\mathrm{M100_{TRF}}$ (Ding and Simon, 2012b; Hertrich et al., 2012), which are themselves localized to posterior auditory cortex (Lütkenhöner and Steinsträter, 1998; Ding and Simon, 2012b). The sensitivity to stimulus contrast has been seen for the MEG auditory steady-state response to 40 Hz amplitude modulations, which also has short latency and localizes to core auditory cortex (Ross et al., 2000). The auditory steady-state response is substantially weakened by a reduction of the stimulus modulation depth (Ross et al., 2000) or an increase of the level of background noise, regardless of the subjects' attentional state (Okamoto et al., 2011). These MEG results are also consistent with animal studies, which demonstrate that neurons in core auditory cortex show contrast gain control but are still sensitive to the modulation depth of the stimulus (Malone et al., 2010; Rabinowitz et al., 2011).

In contrast, almost complete contrast gain control is seen in the long latency $\mathrm{M100_{TRF}}$ component, localized to posterior association auditory cortex (Ding and Simon, 2012a, 2012b). When the subjects actively listen to noise-corrupted speech, the amplitude of the $\mathrm{M100_{TRF}}$ remains unaffected for all SNRs >−9 dB. Similarly, for subjects engaged in a syllable discrimination task, the EEG N1 response to isolated syllables (latency near 100 ms) is also stable to background noise, at least for positive SNRs (Whiting et al., 1998; Kaplan-Neeman et al., 2006). This robustness, however, is not observed during passive listening and therefore may require attention. For example, the EEG N1 response to isolated syllables (Cunningham et al., 2001) or pure tones (Billings et al., 2009) is significantly weakened by background noise during passive listening. Similarly, the auditory steady-state response evoked by slow amplitude modulations (e.g., at 4 Hz), which has latency near 100 ms, also diminishes when the stimulus modulation depth decreases, during passive listening (Rees et al., 1986). In summary, neural adaptation to the dynamic range of stimulus enhances along the ascending auditory pathway, even from the shorter latency (~50 ms) response from core auditory cortex to the longer latency (>100 ms) response from association auditory cortex.

Encoding of slow temporal modulations and long-term integration

As the SNR decreases, the very low-frequency (<4 Hz) neural responses remain robust >−6 dB, whereas the higher frequency (4–8 Hz) neural responses are continuously degraded (Fig. 4*A*). This suggests that, in noisy environments, occurrences of stressed syllables, reflected by very slow (< 4 Hz) temporal modulations (Greenberg, 1999), are more reliably encoded in cortex than faster linguistic structures, such as unstressed syllables and phonemes. This selective neural encoding of very slow temporal modulations can facilitate speech recognition in noise as the very slow temporal modulations of speech are less distorted by noise (Fig. 6*B*). More importantly, psychoacoustic and modeling studies have suggested that intelligibility of speech relies more on the very slow modulations in the presence of stationary noise (Füllgrabe et al., 2009; Jorgensen and Dau, 2011).

The stable neural encoding in low (1–4 Hz) but not higher (4–8 Hz) frequency ranges may be related to the intrinsic properties of cortical neural circuits because δ (1–4 Hz) and *θ* (4–8 Hz) have been classified as distinct frequency bands for cortical oscillations. The current results are consistent with the hypothesis that δ band activity is more strongly related to cognitive control of auditory processing whereas *θ* band activity is more closely tied to the physical properties of the sensory stimulus (Schroeder et al., 2008; Schroeder and Lakatos, 2009).

The robust neural synchronization to slow, but not fast, rhythms of speech reflects a change in the modulation transfer function (i.e., cortical sensitivity to temporal modulations at different modulation rates). As SNR decreases, the cutoff frequency of the stimulus spectrum increases while the cutoff frequency of the response spectrum decreases. This clearly indicates a loss of sensitivity to the noise-corrupted higher modulation rates (e.g., 5–10 Hz). Similar changes of modulation sensitivity have been shown in individual neurons from anesthetized animals: Neurons in midbrain of both songbirds and gerbils are more sensitive to higher modulation rates when stimulated by animal vocalizations than when stimulated by noise (Woolley et al., 2006; Lesica and Grothe, 2008). In mammalian auditory cortex, the temporal sensitivity of neurons can be further modulated by top-down attention (Fritz et al., 2007).

Parsing of continuous speech and intelligibility

The neural synchronization to slow temporal modulations <4 Hz is more robust to noise than speech intelligibility and therefore is more likely to reflect the perception of the prosody of speech, which is more robust than the recognition of phonemes (Woodfield and Akeroyd, 2010). Even though the neural synchronization to slow modulations is more robust to noise than intelligibility and therefore not a good indicator of how intelligibility is affected by noise, it does predict how well individual subjects recognize speech in noise (Fig. 2*C*). This indicates that, for the same noisy speech stimulus, subjects with a more faithful auditory cortical representation of speech can understand it better (for a discussion at the brainstem level, see Ruggles et al., 2012). This correlation is especially informative because speech recognition is a very complex process, involving extensive cortical areas beyond auditory cortex (Scott et al., 2004). The relatively high correlation (~0.8, between auditory encoding accuracy and speech scores) indicates, for speech recognition in noise, a major bottleneck lies in auditory processing (i.e., ability to extract speech information from the acoustic background).

In conclusion, this study demonstrates that the cortical entrainment to the very slow rhythm of speech (< 4 Hz) is robust to background noise. This indicates that, although the acoustic features of speech are severely corrupted by background noise, a nearly noise-invariant neural representation of the prosody of speech is formed in auditory cortex, through complex neural computations that include contrast gain control and adaptive neural encoding of temporal modulations. Furthermore, the precision of cortical entrainment, which is variable over listeners, provides a neural correlate of how well each listener can understand speech in noise.

References

Anderson S, Skoe E, Chandrasekaran B, Kraus N (2010) Neural timing is linked to speech perception in noise. J Neurosci 30:4922–4926. CrossRef Medline

Bar-Yosef O, Nelken I (2007) The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. Front Comput Neurosci 1:3. CrossRef Medline

Billings CJ, Tremblay KL, Stecker GC, Tolin WM (2009) Human evoked cortical activity to signal-to-noise ratio and absolute signal level. Hearing Res 254:15–24. CrossRef Medline

Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 109:1101–1109. CrossRef Medline

Cunningham J, Nicol T, Zecker SG, Bradlow A, Kraus N (2001) Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement. Clin Neurophysiol 112:758–767. CrossRef Medline

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectrotemporal receptive fields with natural stimuli. Network 18:191–212. CrossRef Medline

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. J Neurosci 8:1684–1689. CrossRef Medline

de Cheveigné A, Simon JZ (2008) Denoising based on spatial filtering. J Neurosci Methods 171:331–339. CrossRef Medline

Delgutte B (1980) Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J Acoust Soc Am 68:843–857. CrossRef Medline

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

Ding N, Simon JZ (2012b) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859. CrossRef Medline

Escabí MA, Miller LM, Read HL, Schreiner CE (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J Neurosci 23:11489–11504. Medline

Festen JM, Plomp R (1990) Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am 88:1725–1736. CrossRef Medline

Fishman YI, Steinschneider M (2010) Formation of auditory streams. In: The oxford handbook of auditory science: the auditory brain (Rees A, Palmer A, eds), pp 215–245: New York: Oxford UP.

Fritz JB, Elhilali M, David SV, Shamma SA (2007) Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? Hearing Res 229:186–203. CrossRef Medline

Füllgrabe C, Stone MA, Moore BC (2009) Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task. J Acoust Soc Am 125:1277–1280. CrossRef Medline

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15: 511–517. CrossRef Medline

Greenberg S (1999) Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation. Speech Commun 29:159–176. CrossRef

Greenberg S, Carvey H, Hitchcock L, Chang S (2003) Temporal properties of spontaneous speech: a syllable-centric perspective. J Phonetics 31:465–485. CrossRef

Griffiths TD, Warren JD (2004) What is an auditory object? Nat Rev Neurosci 5:887–892. CrossRef Medline

Hertrich I, Dietrich S, Trouvain J, Moos A, Ackermann H (2012) Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. Psychophysiology 49:322–334. CrossRef Medline

Jørgensen S, Dau T (2011) Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J Acoust Soc Am 130:1475–1487. CrossRef Medline

Kaplan-Neeman R, Kishon-Rabin L, Henkin Y, Muchnik C (2006) Identification of syllables in noise: electrophysiological and behavioral correlates. J Acoust Soc Am 120:926–933. CrossRef Medline

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party." J Neurosci 30:620–628. CrossRef

Lesica NA, Grothe B (2008) Efficient temporal processing of naturalistic sounds. PLoS ONE 3:e1655. CrossRef Medline

Lütkenhöner B, Steinsträter O (1998) High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol Neurootol 3:191–213. CrossRef Medline

Malone BJ, Scott BH, Semple MN (2010) Temporal codes for amplitude contrast in auditory cortex. J Neurosci 30:767–784. CrossRef Medline

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

Mosher JC, Baillet S, Leahy RM (2003) Equivalence of linear approaches in bioelectromagnetic inverse solutions. Paper presented at IEEE Workshop on Statistical Signal Processing, September 28–October 1, St. Louis.

Nelken I, Rotman Y, Bar Yosef O (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397:154–157. CrossRef Medline

Okamoto H, Stracke H, Bermudez P, Pantev C (2011) Sound processing hierarchy within human auditory cortex. J Cogn Neurosci 23:1855–1863. CrossRef Medline

Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–113. CrossRef Medline

Rabinowitz NC, Willmore BD, Schnupp JW, King AJ (2011) Contrast gain control in auditory cortex. Neuron 70:1178–1191. CrossRef Medline

Rees A, Green G, Kay R (1986) Steady-state evoked responses to sinusoidally amplitude-modulated sounds recorded in man. Hearing Res 23:123–133. CrossRef Medline

Robinson BL, McAlpine D (2009) Gain control mechanisms in the auditory pathway. Curr Opin Neurobiol 19:402–407. CrossRef Medline

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336:367–373. CrossRef Medline

Ross B, Borgmann C, Draganova R, Roberts LE, Pantev C (2000) A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. J Acoust Soc Am 108:679–691. CrossRef Medline

Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2012) Why middle-aged listeners have trouble hearing in everyday settings. Curr Biol 22:1858. CrossRef Medline

Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32:9–18. CrossRef Medline

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. Trends Cogn Sci 12:106–113. CrossRef Medline

Scott SK, Rosen S, Wickham L, Wise RJ (2004) A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. J Acoust Soc Am 115:813–821. CrossRef Medline

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34:114–123. CrossRef Medline

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

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182–186. CrossRef Medline

Stone MA, Füllgrabe C, Mackinnon RC, Moore BC (2011) The importance for speech intelligibility of random fluctuations in "steady" background noise. J Acoust Soc Am 130:2874–2881. CrossRef Medline

Watkins PV, Barbour DL (2008) Specialized neuronal adaptation for preserving input sensitivity. Nat Neurosci 11:1259–1261. CrossRef Medline

Wen B, Wang GI, Dean I, Delgutte B (2009) Dynamic range adaptation to sound level statistics in the auditory nerve. J Neurosci 29:13797–13808. CrossRef Medline

Whiting KA, Martin BA, Stapells DR (1998) The effects of broadband noise masking on cortical event-related potentials to speech sounds /ba/ and /da/. Ear Hear 19:218–231. CrossRef Medline

Woodfield A, Akeroyd MA (2010) The role of segmentation difficulties in speech-in-speech understanding in older and hearing-impaired adults. J Acoust Soc Am 128:EL26–EL31. CrossRef Medline

Woolley SM, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J Neurosci 26:2499–2512. CrossRef Medline

Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. IEEE Trans Information Theory 38:824–839. CrossRef

Zilany MS, Bruce IC, Nelson PC, Carney LH (2009) A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. J Acoust Soc Am 126:2390–2412. CrossRef Medline