NeuroImage 88 (2014) 41–46

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure

Nai Ding [a,b], Monita Chatterjee [c], Jonathan Z. Simon [a,d,e,*]

[a] Department of Electrical and Computer Engineering, University of Maryland, College Park, College Park, MD 20742, USA
[b] Department of Psychology, New York University, New York, NY 10003, USA
[c] Boys Town National Research Hospital, Omaha, NE 68131, USA
[d] Department of Biology, University of Maryland, College Park, College Park, MD 20742, USA
[e] Institute for Systems Research, University of Maryland, College Park, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Accepted 27 October 2013
Available online 2 November 2013

Keywords:
Envelope entrainment
Auditory cortex
Auditory scene analysis
MEG

ABSTRACT

Speech recognition is robust to background noise. One underlying neural mechanism is that the auditory system segregates speech from the listening background and encodes it reliably. Such robust internal representation has been demonstrated in auditory cortex by neural activity entrained to the temporal envelope of speech. A paradox, however, then arises, as the spectro-temporal fine structure rather than the temporal envelope is known to be the major cue to segregate target speech from background noise. Does the reliable cortical entrainment in fact reflect a robust internal "synthesis" of the attended speech stream rather than direct tracking of the acoustic envelope? Here, we test this hypothesis by degrading the spectro-temporal fine structure while preserving the temporal envelope using vocoders. Magnetoencephalography (MEG) recordings reveal that cortical entrainment to vocoded speech is severely degraded by background noise, in contrast to the robust entrainment to natural speech. Furthermore, cortical entrainment in the delta-band (1–4 Hz) predicts the speech recognition score at the level of individual listeners. These results demonstrate that reliable cortical entrainment to speech relies on the spectro-temporal fine structure, and suggest that cortical entrainment to the speech envelope is not merely a representation of the speech envelope but a coherent representation of multiscale spectro-temporal features that are synchronized to the syllabic and phrasal rhythms of speech.

Introduction

Normal hearing listeners exhibit a surprising ability to understand speech in noisy acoustic environments, even in the absence of visual cues. A number of studies have suggested that the target speech and the listening background are separated in auditory cortex (Ding and Simon, 2012a; Zion Golumbic et al., 2013; Horton et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012; Power et al., 2012). In particular, when a listener attends to a speech stream, auditory cortical activity is reliably entrained to the temporal envelope of that stream, regardless of the listening background. This reliable neural representation of the speech envelope, i.e. slow temporal modulations below 16 Hz, is a key candidate mechanism underlying the reliable recognition of speech, since the temporal envelopes carry important cues for speech recognition (Shannon et al., 1995). It remains mysterious, however, how such reliable cortical entrainment to the speech envelope is achieved, since envelope is not an effective cue for segregation of speech from noise (Friesen et al., 2001).

Moreover, even the nature of cortical entrainment to the speech envelope is heavily debated, especially about whether it encodes the temporal envelope per se or instead other speech features that are correlated with the speech envelope (Obleser et al., 2012; Peelle et al., 2013). Many speech features, including pitch and spatial cues, are temporally coherent and correlated with the temporal envelope (Shamma et al., 2011). Therefore it has been proposed that the envelope entrainment in fact reflects a collective neural representation of multiple speech features that are synchronized to the syllabic and phrasal rhythm of speech (Ding and Simon, 2012a). Because of the collective nature of this representation, it has been suggested as a representation of speech as a whole auditory object.

If envelope entrainment indeed reflects an object-level, collective representation of speech features, reliable envelope entrainment in complex auditory scenes is likely to involve an analysis-by-synthesis process (Poeppel et al., 2008; Shamma et al., 2011; Shinn-Cunningham, 2008): In such a process, multiple features of a complex auditory scene are extracted subcortically in the analysis phase and then, based on speech segregation cues such as pitch, features belonging to the same speech stream are grouped into an auditory object in the synthesis phase. In

* Corresponding author at: Department of Biology, University of Maryland, College Park, College Park, MD 20742, USA. Fax: +1 301 314 9281.
E-mail addresses: gahding@gmail.com (N. Ding), monita.chatterjee@boystown.org (M. Chatterjee), jzsimon@umd.edu (J.Z. Simon).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.10.054

contrast, if envelope entrainment involves only direct neural processing of the envelope, its robustness to noise may arise from more basic processes such as contrast gain control (Ding and Simon, 2013; Rabinowitz et al., 2011).

In this study, we investigate whether noise-robust cortical entrainment to the speech envelope involves merely envelope processing or instead reflects an analysis-by-synthesis process that includes the processing of spectro-temporal fine structure and reflects envelope properties of the re-synthesized auditory object. Here, the spectro-temporal fine structure refers to the acoustic information not included in the broadband envelope of speech (<16 Hz), including, for example, the acoustic cues responsible for the pitch and formant structure of speech. We degrade the spectro-temporal fine structure of speech or speech–noise mixtures using noise vocoders and investigate whether vocoded stimuli are cortically represented differently from natural speech using MEG. If cortical entrainment only depends on the temporal envelope, it will not be affected by degradation of the spectro-temporal fine structure, even in a noisy listening environment. In contrast, if reliable cortical entrainment to speech requires an analysis-by-synthesis process that relies on the spectro-temporal fine structure, it should be severely degraded for vocoded speech.

Materials & methods

Subjects

Twelve normal hearing, right-handed (Oldfield, 1971) young adults (6 females), all between 19 and 32 years old (23 years old on average) participated in the experiment. Subjects were paid, and the experimental procedures were approved by the University of Maryland institutional review board. Written informed consent form was obtained before the experiment.

Stimuli

The stimuli were selected from of a narration of the story *Alice's Adventures in Wonderland* (Chapter One, http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-4/). The sound recording was low-pass filtered below 4 kHz and divided into twelve 50-second duration segments, after long speaker pauses (>300 ms) were shortened to 300 ms. All sound stimuli were presented binaurally (diotically). Six types of stimuli were created (2 noise levels × 3 vocoding conditions).

Background noise

Half of the speech segments (N = 6) were presented in a quiet listening environment (no noise added in), while the other half were mixed with spectrally matched stationary noise generated using a 12th-order linear predictive model estimated from the speech recording. The intensity ratio between speech and noise was fixed at 3 dB, measured by RMS.

Noise vocoding

Each stimulus is either noise vocoded (through a 4-channel or 8-channel vocoder) or unprocessed. The noise vocoder filters the stimulus, either speech in quiet or speech in noise, into 4 or 8 frequency channels between 123 and 3951 Hz using a 4th order Butterworth filter. All frequency channels are evenly distributed in the Cam scale (Glasberg and Moore, 1990; Qin and Oxenham, 2003). In each frequency band, the envelope of the stimulus, either speech or a speech–noise mixture, is extracted by taking the absolute value of the Hilbert Transform, low-pass filtering below 160 Hz using a 4th order Butterworth filter, and then half-wave rectifying the filtered signal. The extracted envelope is used to modulate white noise filtered into the same frequency band from which the envelope was derived. The envelope-modulated-noises are then summed over frequency bands to create the noise-vocoded stimulus. The RMS intensity of the noise-vocoded stimulus is adjusted to match that of the unprocessed stimulus.

Stimulus characterization

The auditory spectrogram of the stimulus was calculated using a sub-cortical auditory model (Yang et al., 1992) and expressed in a logarithmic amplitude scale. The frequency by time auditory spectrogram has 128 logarithmically spaced frequency channels and a 10-ms resolution in time. The broadband temporal envelope of the stimulus was extracted by summing the auditory spectrogram over frequency.

Procedure

The stimuli were presented in two orders, each to half of the subjects. In either order, the story continued naturally between stimuli and was repeated twice after the first presentation (3 trials in total). In the progressive order, the first two speech segments were natural speech presented in quiet, followed by 8-band vocoded speech in quiet and then 4-band vocoded speech in quiet. Then, natural speech in noise, 8-band vocoded speech in noise, and 4-band vocoded speech in noise were presented sequentially. To control for the effect of presentation order, we also created a random order condition, in which each acoustic manipulation (e.g. vocoding or background noise) was assigned randomly to a segment for each subject. The two presentation orders did not result in any difference in speech intelligibility or neural synchronization spectrum and were therefore not distinguished in the following analysis.

The subjects were asked to listen to the story and keep their eyes closed. Questions about the story were asked after each 50-second duration stimulus to ensure subjects' attention. The subjects were also asked to rate the percent of words they understood after the first presentation of each stimulus (on a scale of 0% (not intelligible) to 100% (fully intelligible)). The grand averaged subjectively rated intelligibility is highly correlated with the grand averaged percent of questions correctly answered ($R = 0.96$). Before the experiment, the subjects listened to 100 repetitions of a 500-Hz tone and the responses were used to extract the M100 response, a salient MEG response localized to auditory cortex (Lütkenhöner and Steinsträter, 1998).

The magnetic field generated by cortical activity was recorded using a 157-channel whole-head MEG system (KIT, Kanazawa, Japan). The signal was sampled at 1 kHz and was filtered by a 200-Hz lowpass filter and a notch filter at 60 Hz online. Environmental noise was further removed using TS-PCA (de Cheveigné and Simon, 2007). The whole-head MEG recording was used for analysis unless otherwise specified. When the two hemispheres were analyzed separately, hemisphere-specific responses were extracted using 55 sensors located above each hemisphere. More details of the recording procedure are as described in Ding and Simon (2012a).

Inter-trial correlation analysis

The phase locking of a neural response was evaluated by the inter-trial correlation of the neural response in narrow frequency bands (2-Hz wide) (Ding and Simon, 2013; Zion Golumbic et al., 2013). The inter-trial correlation is the Pearson correlation coefficient between two trials of the neural responses to the same stimulus (averaged over all possible combinations of two trials). It measures the reliability of the neural response when the same stimulus repeats, and reflects the strength of phase-locked neural activity. The major phase-locked component of the MEG response was extracted using a blind source separation method, Denoising Source Separation (DSS) (de Cheveigné and Simon, 2008). The first DSS component was used for this analysis.

Temporal response function

The response from each MEG sensor is modeled as the speech envelope convolved with a temporal response function (TRF), which characterizes the cortical response evoked by a unit power increase of the stimulus (Ding and Simon, 2012b). The TRF is derived by summing a spectro-temporal response function (STRF) over frequency.

The STRF is estimated using boosting with 10-fold cross validation (David et al., 2007). For computational efficiency, the 157 MEG sensors were reduced to 10 DSS components (de Cheveigné and Simon, 2008). The STRFs separately estimated for the 10 DSS components were converted back to the sensor space for further analysis (Ding and Simon, 2012a).

The two major peaks of the TRF have latencies near 50 ms and 100 ms and they are referred to as the $M50_{TRF}$ and the $M100_{TRF}$. The $M50_{TRF}$ is extracted as the response peak, identified by the root mean square (RMS) of the MEG response over sensors, between 20 and 80 ms, while the $M100_{TRF}$ is extracted as the peak between 90 and 160 ms.

When deriving the STRF, the auditory spectrogram (Yang et al., 1992) of clean speech is always used, even when modeling the neural response to noisy speech. Previous studies have shown that the spectrogram of clean speech models the MEG response slightly better than the spectrogram of the actual noisy stimulus (Ding and Simon, 2013). More importantly, since the background noise is stationary, the shape of the spectrogram of the noisy stimulus closely resembles that of clean speech but the dynamic range, or contrast, of the spectrogram is much smaller for the noisy stimulus. Therefore, it is the gain, rather than the shape of the TRF, that depends strongly on which spectrogram is used. Here, by using the original speech spectrogram, the effect of noise on the TRF amplitude depends on how the neural response amplitude changes with noise, rather than on how the stimulus dynamic range changes with noise.

Results

MEG responses were recorded from subjects listening to a narrated story presented either in quiet or in spectrally matched stationary noise (3 dB SNR). The speech stimuli were presented either without additional processing, referred to as natural speech, or after being processed by a noise vocoder (4-band or 8-band), referred to as vocoded speech. Noise vocoding reduces the spectral resolution of speech, as is demonstrated by the auditory spectrograms of the stimuli (Fig. 1). The temporal envelope (summation of the auditory spectrogram over frequencies), however, is essentially identical before and after noise vocoding ($R > 0.99$ for the stimuli used in this study).

Neural synchronization spectrum

We first characterize how each stimulus synchronizes the neural responses in different frequency bands. The degree of neural synchronization is measured by the inter-trial correlation of the neural recording (Fig. 2). Consistent with previous studies (Ding and Simon, 2012b; Luo and Poeppel, 2007), neural synchronization to speech is observed in the delta (1–4 Hz) and theta (4–8 Hz) bands. A comparison of the neural responses to speech in quiet and speech in noise indicates that the degree of neural synchronization was robust to noise for natural but not for noise-vocoded speech (Fig. 2A). Note that for speech presented in noise, the same noise signal is used across trials. Therefore, the inter-trial correlation reflects the neural phase locking to any available stimulus features, including the background noise. Therefore, the reduced neural phase locking for vocoded speech in noise indicates an overall reduction in the response to the speech–noise mixture.

In a quiet listening environment, as the spectral resolution of the stimulus decreases, neural synchronization below 4 Hz is enhanced ($P < 0.01$, 1-way repeated measures ANOVA) while neural synchronization above 4 Hz is reduced ($P < 0.003$, 1-way repeated measures ANOVA) (Fig. 2BC). In a noisy listening environment, however, the degree of neural synchronization is reduced in both the delta ($P < 0.003$, 1-way repeated measures ANOVA) and theta bands ($P < 10^{-5}$, 1-way repeated measures ANOVA) as the stimulus spectral resolution reduces (Fig. 2B). In this analysis, the two hemispheres are combined. When each hemisphere is analyzed separately, no significant hemispherical lateralization is seen in any of the 6 stimulus conditions for any 2-Hz band between 2 and 8 Hz ($P > 0.17$ and on average 0.37, uncorrected paired t-test).

Predicting individual speech recognition score

The subjectively rated speech recognition score varies strongly across subjects. The individual recognition score significantly correlates with delta-band neural synchronization for 4-band vocoded speech in quiet, and for 8-band vocoded speech both in quiet and in noise ($P < 0.002$, bootstrap). Since there are 6 stimulus conditions, the P-value remains below 0.012 after a Bonferroni correction. The correlation coefficients are 0.66 ± 0.14, 0.55 ± 0.14, and 0.71 ± 0.11 (mean $\pm$ SEM) for these 3 conditions (Fig. 3, from left to right). For 4-band vocoded speech in noise, a weaker correlation is also found ($P < 0.02$, bootstrap; $R = 0.43 \pm 0.20$). For natural speech in quiet and in noise, speech intelligibility reaches ceiling, obscuring any observable correlation between neural synchronization and speech intelligibility. In this correlation analysis, the two hemispheres are combined. If each hemisphere is considered separately, the only

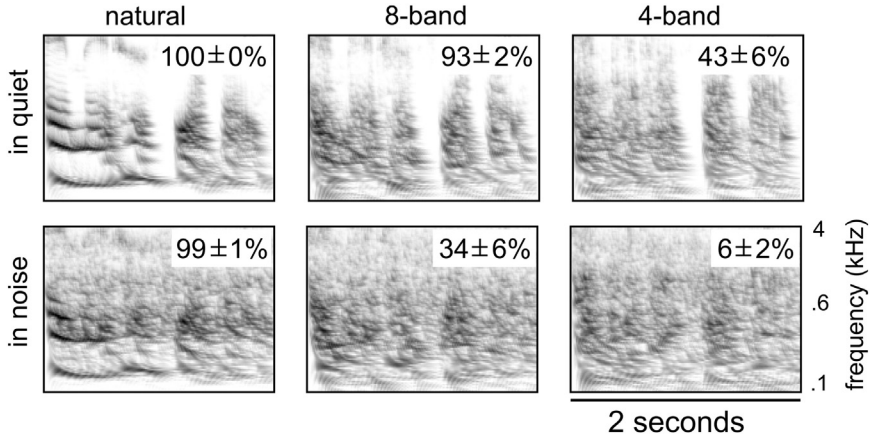

Fig. 1. Examples of the auditory spectrograms of the experimental stimuli (2 noise levels × 3 vocoding conditions). Subjectively rated speech recognition score (mean ± SEM) is labeled in the upper right corner of each spectrogram. Vocoding (8-band or 4-band) degrades the spectro-temporal fine structure of speech, for example, the harmonic structure, but preserves the temporal envelope.

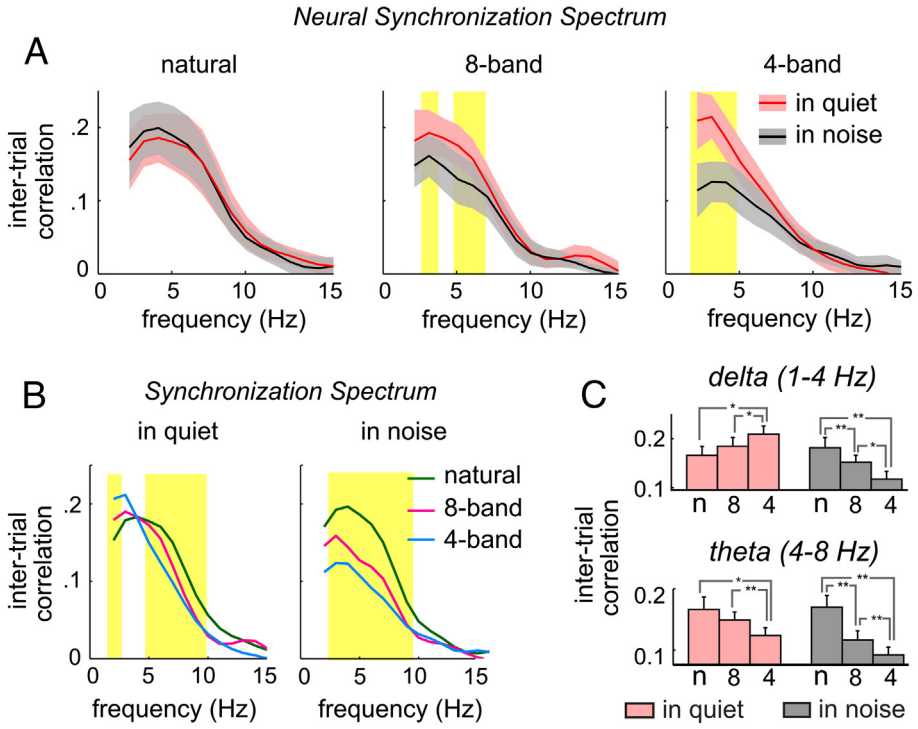

Fig. 2. The inter-trial correlation of the neural response to natural and vocoded speech. (A) The inter-trial correlation grouped by spectro-resolution. The background noise decreases the inter-trial correlation of the neural response for vocoded speech but not for clean speech. Frequency regions where the response is significantly affected by noise ($P < 0.001$, 1-way ANOVA) are shaded in yellow. (B) The inter-trial correlation grouped by the noise level. Noise vocoding affects the phase-locking spectrum differentially for speech in quiet and speech in noise. Frequency regions where the response is significantly affected by vocoding ($P < 0.001$, 1-way ANOVA) are shaded in yellow. (C) The inter-trial correlation averaged over the delta (1–4 Hz) and theta (4–8 Hz) bands. The error bar represents one SEM. In quiet, delta-band inter-trial correlation increases with reduced spectral resolution, while theta-band inter-trial correlation decreases. For speech in noise, the inter-trial correlation decreases with decreasing spectral resolution for both bands. $^{*}P < 0.01$, $^{**}P < 0.001$, (paired t-test).

significant correlation is for the 8-band vocoded speech in noise, in the left hemisphere ($P < 0.002$, bootstrap). This result indicates that the two hemispheres encode speech similarly and therefore integrating measurements across hemispheres increases the statistical power of the correlation analysis. The correlation between neural synchronization and speech intelligibility is only observed in the delta band but not in the theta band in any condition.

Temporal response function

The neural synchronization analysis characterizes the response reliability over trials, while in the following we further investigate how the neural response follows the speech envelope using a temporal response function (TRF). The TRF can be interpreted by the neural response evoked by a broadband power increase of the stimulus (Ding and Simon, 2012b). The RMS of the TRFs from all MEG sensors is shown in Fig. 4. The amplitude of the TRF is dimensionless, and is normalized by the maximal amplitude of the TRF for natural speech in quiet. The TRF in quiet shows two early peaks near 50 ms and 100 ms and these two peaks are referred to as the $M50_{TRF}$ and $M100_{TRF}$ respectively (Ding and Simon, 2013).

The early response component $M50_{TRF}$ is sensitive to noise while the late response component $M100_{TRF}$ is not. Specifically, the $M50_{TRF}$ amplitude is significantly reduced by noise ($P < 0.001$, $F(1, 71) = 41.26$, SNR × spectral resolution 2-way repeated measures ANOVA) and there is an interaction between the influence of noise and the influence of spectral resolution ($P = 0.033$ with Geisser-Greenhouse corrections, $F(2, 71) = 5.34$). To investigate the interaction between noise and spectral resolution, two separate ANOVA tests are applied to the $M50_{TRF}$ amplitude in quiet and in noise, with spectral resolution as the analysis factor. In quiet, the M50 amplitude increases with reduced spectral resolution ($P = 0.006$, $F(2, 33) = 6.03$, 1-way repeated measures ANOVA). In noise, the $M50_{TRF}$ amplitude is weak and is not significantly changed when the stimulus spectral resolution reduces. The $M100_{TRF}$ amplitude is not significantly affected by noise or resolution.

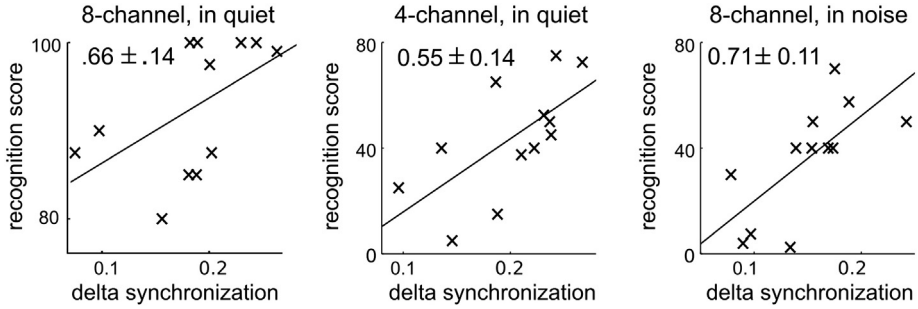

Fig. 3. Delta-band neural synchronization correlates with the speech recognition score of individual listeners. Each cross shows the data from a listener and the solid line is the regression line. The correlation coefficient between delta-band inter-trial correlation and the speech score is shown at the upper left corner of each plot (mean ± SEM).

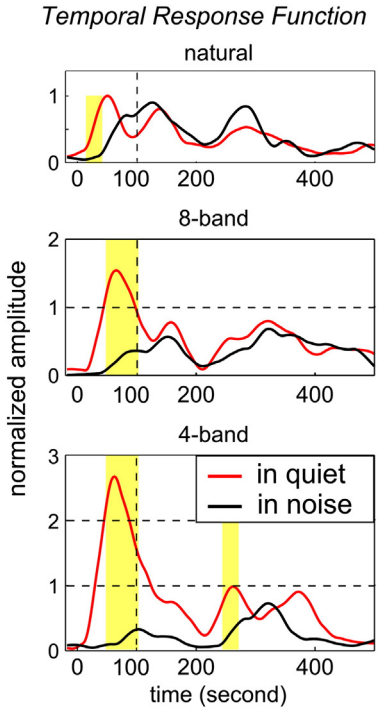

Fig. 4. Properties of the temporal response function (TRF). (A) The RMS of the TRF over MEG sensors, which can be interpreted as the total energy of the neural response evoked by a unit power increase of the sound stimulus. Time intervals where the TRF amplitude is significantly modulated by background noise ($P < 0.001$) are shaded in yellow. The early peak of the TRF near 50 ms, i.e. $M50_{TRF}$, is most strongly affected by background noise and stimulus spectral resolution. Its amplitude decreases in the presence of background noise and increases when the stimulus spectral resolution decreases in a quiet listening environment.

Discussion

This study demonstrates that although the cortical entrainment to natural speech is robust to noise the cortical entrainment to vocoded speech is not. This phenomenon cannot be explained by passive envelope tracking mechanisms since noise vocoding does not directly affect the stimulus envelope to which that cortical activity is entrained. Instead, the results illustrate that the spectro-temporal fine structure, which is degraded for noise-vocoded speech, is critical to segregating speech from noise and to constructing an object-level neural representation of speech that is robust to the listening background.

Object-based vs. stimulus-based representation

Only after simultaneous auditory objects are neurally segregated, can each of them be represented and processed independent of each other (Ding and Simon, 2012a). On the other hand, if the auditory scene is represented as a whole in auditory cortex, the cortical representation will be affected by every component in the same auditory scene. In this study, we found that the neural representation of natural speech is largely invariant to a moderate amount of background noise, i.e. at 3 dB SNR, indicating that natural speech is neurally segregated from noise and is represented as an individual auditory object. This result is consistent with a previous study where the neural representation of speech is found to be largely independent of background noise until the significantly worse case of −3 dB SNR (Ding and Simon, 2013). In contrast, the neural representation of vocoded speech is significantly degraded by noise. This encoding interference between noise and speech suggests that the vocoded speech noise mixture is not neurally segregated in auditory cortex.

The role of contrast gain control

Stationary noise significantly reduces the intensity contrast of speech but does not strongly affect the shape of the temporal envelope of speech (Fig. 1). As a result, the robust neural representation of natural speech may be accounted for by contrast gain control (Ding and Simon, 2013), a relatively passive mechanism that can be observed in anesthetized animals (Dean et al., 2005; Rabinowitz et al., 2011). Nevertheless, although contrast and intensity gain control surely play an important role in maintaining the noise robust representation of speech, they cannot explain the phenomenon observed in this study without assuming the prior neural segregation of speech and noise. First, background noise reduces the intensity contrast of natural speech and vocoded speech in the same way, but a noise-robust cortical representation is only observed for natural speech. Second, even in auditory cortex, contrast gain control is incomplete, in the sense that even though the neural gain changes, the cortical response is still affected by the stimulus intensity contrast (Rabinowitz et al., 2011).

Robust cortical entrainment: An analysis-by-synthesis approach

An analysis-by-synthesis approach is ubiquitously adopted in sensory systems (Poeppel et al., 2008; Yuille and Kersten, 2006). In a primary analysis stage, the sensory system breaks up the sensory input into fundamental features, e.g. edges are encoded in the visual system and spectro-temporal features in the auditory system. After this stage, however, a synthesis stage is necessary to reconstruct the sensory experience, usually with top-down modulations that contribute pertinent a priori information about the observer's world. Such mechanisms are likely to be useful in attenuating the effects of unwanted noise.

The sensitivity to the spectro-temporal fine structure indicates that robust cortical entrainment to speech is the consequence of the analysis-by-synthesis process rather than just bottom-up envelope tracking. In the sub-cortical auditory system, acoustic cues that are important for sound source segregation, such as pitch and binaural cues, are extracted (Nelken, 2008). This decomposition process can be viewed as an analysis stage. In order to achieve speech recognition or auditory perception in general, however, features belonging to the same speech stream need to be bound or re-synthesized into an auditory object (Shinn-Cunningham, 2008). In speech, multiple acoustic features are temporally coupled and the spectro-temporal fine structure is modulated by the temporal envelope (Shamma et al., 2011; Sheft, 2007). Therefore, in this synthesis stage, sound segregation cues play a guiding role: The auditory system is proposed to group features based on their temporal coherence with the sound segregation cues (Shamma et al., 2011). As a consequence of this temporal coherence based grouping, features belonging to the attended speech stream are recovered from a complex auditory scene and appear as neural activity entrained to the temporal envelope of the attended speech. When sound segregation cues such as the spectro-temporal fine structure are degraded, features extracted from a complex auditory scene can no longer be selectively grouped into a representation specific to the attended speech stream. Therefore, cortical entrainment is degraded.

Influence of spectral resolution on speech encoding

As the spectral resolution of speech is reduced, speech intelligibility decreases mildly in a quiet listening environment but severely in noisy environments (Friesen et al., 2001). The same trend is seen in theta- but not delta-band cortical synchronization. In a quiet environment, as the stimulus spectral resolution decreases, theta-band synchronization is moderately reduced, consistent with previous studies (Luo and Poeppel, 2007; Peelle et al., 2013), while delta-band synchronization is enhanced. It is possible that the reduction in theta-band activity reflect an impairment of neural processing of syllabic-level speech features (Giraud and Poeppel, 2012; Peelle et al., 2013), while the

enhancement in delta-band activity, hypothesized as an instrument of top-down attention (Schroeder and Lakatos, 2009), may reflect increased listening effort. Both the delta- and theta-band synchronizations in auditory cortex are likely to be modulated by higher level cortical areas involved in language processing (Obleser and Weisz, 2012; Scott et al., 2006). Furthermore, when the spectro-temporal resolution is reduced in the time domain, the $M50_{TRF}$ is enhanced. This is likely to be related to the observations that the $M50_{TRF}$ is stronger at the onset of a noise burst than at the onset of a tone (Chait et al., 2004) and that the onset response to vocoded speech is stronger than the onset response to natural speech (Obleser and Kotz, 2010).

Delta band synchronization and speech intelligibility

Cortical entrainment to speech is generally observed in the delta and theta bands. In this study and other studies using long (>10 s) continuous speech stimuli, delta-band entrainment dominates the measured neural activity (e.g. Ding and Simon, 2012a, 2013; Zion Golumbic et al., 2013). For studies using isolated sentences (<5 s in duration), however, delta-band entrainment is much weaker than theta-band entrainment (e.g. Howard and Poeppel, 2010; Luo and Poeppel, 2007; Peelle et al., 2013). Therefore, it is likely that delta-band entrainment requires the longer time scale contextual information of running speech.

Here, we observed that delta-band synchronization correlates with listeners' speech recognition scores for vocoded speech, and a previous study found a similar correlation for speech embedded in strong noise (Ding and Simon, 2013). It is possible that delta-band synchronization to speech is a signature of the auditory cortical representation subserving subsequent language processing (Giraud and Poeppel, 2012; Schroeder and Lakatos, 2009; Schroeder et al., 2008). Alternatively, it is possible that speech intelligibility is required for delta-band synchronization to occur. This possibility, however, is not well supported since strong neural synchronization has been seen to reversed speech (Howard and Poeppel, 2010) and amplitude/frequency modulated tones (Henry and Obleser, 2012; Wang et al., 2012).

The correlation between neural synchronization and individual speech recognition score is not found in the theta band, consistent with previous studies (Peelle et al., 2013). One possible reason is that theta-band synchronization is relatively weak compared with delta-band synchronization for discourse level speech stimuli, and the lower signal to noise ratio of theta activity makes it less reliably measured at an individual subject level. Alternatively, it is also possible that theta band activity faithfully reflects properties of the stimulus but not individual differences in neural processing (see also Schroeder et al., 2008).

In summary, the spectro-temporal fine structure is required to maintain noise-robust cortical entrainment to the speech envelope. These results demonstrate that envelope entrainment in auditory cortex is not just a neural representation of the speech envelope per se but instead is likely to be a collective, object-level neural representation that is achieved by an analysis-by-synthesis approach. Furthermore, since degraded ability to separate simultaneous auditory objects is common for hearing impaired listeners (Shinn-Cunningham and Best, 2008), the results here are indicative of the cortical processing in impaired auditory systems.

Acknowledgment

We thank NIH grants R01 DC 008342 to J.Z.S. and R01 DC 004786 to M.C. for support.

References

Chait, M., Simon, J.Z., Poeppel, D., 2004. Auditory M50 and M100 responses to broadband noise: functional implications. Neuroreport 15, 2455–2458.
David, S.V., Mesgarani, N., Shamma, S.A., 2007. Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw. Comput. Neural Syst. 18, 191–212.
de Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift PCA. J. Neurosci. Methods 165, 297–305.
de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. J. Neurosci. Methods 171, 331–339.
Dean, I., Harper, N.S., McAlpine, D., 2005. neural population coding of sound level adapts to stimulus statistics. Nat. Neurosci. 8, 1684–1689.
Ding, N., Simon, J.Z., 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. U. S. A. 109, 11854–11859.
Ding, N., Simon, J.Z., 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89.
Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33, 5728–5735.
Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110, 1150–1163.
Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.
Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47, 103–138.
Henry, M.J., Obleser, J., 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc. Natl. Acad. Sci. 109, 20095–20100.
Horton, C., D'Zmura, M., Srinivasan, R., 2013. Suppression of competing speech through entrainment of cortical oscillations. J. Neurophys. 109, 3082–3093.
Howard, M.F., Poeppel, D., 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J. Neurophysiol. 104, 2500–2511.
Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. 30, 620–628.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.
Lütkenhöner, B., Steinsträter, O., 1998. High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol. Neuro Otol. 3, 191–213.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.
Nelken, I., 2008. Processing of complex sounds in the auditory system. Curr. Opin. Neurobiol. 18, 413–417.
Obleser, J., Kotz, S.A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb. Cortex 20, 633–640.
Obleser, J., Weisz, N., 2012. Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. Cereb. Cortex 22, 2466–2477.
Obleser, J., Herrmann, B., Henry, M.J., 2012. Neural oscillations in speech: don't be enslaved by the envelope. Front. Hum. Neurosci. 6.
Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension cerebral cortex. Cereb. Cortex. http://dx.doi.org/10.1093/cercor/bhs1118.
Poeppel, D., Idsardi, W.J., Vv, Wassenhove, 2008. Speech perception at the interface of neurobiology and linguistics. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 363, 1071–1086.
Power, A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., Lalor, E.C., 2012. At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. 35, 1497–1503.
Qin, M.K., Oxenham, A.J., 2003. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am. 114, 446–454.
Rabinowitz, N.C., Willmore, B.D.B., Schnupp, J.W.H., King, A.J., 2011. Contrast gain control in auditory cortex. Neuron 70, 1178–1191.
Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32, 9–18.
Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. 12, 106–113.
Scott, S.K., Rosen, S., Lang, H., Wise, R.J., 2006. Neural correlates of intelligibility in speech investigated with noise vocoded speech — a positron emission tomography study. J. Acous. Soc. Am. 120, 1075–1083.
Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34, 114–123.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.
Sheft, S., 2007. Envelope processing and sound-source perception. In: Yost, W.A., Popper, A.N., Fay, R.R. (Eds.), Auditory Perception of Sound Sources. Springer, New York.
Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. 12, 182–186.
Shinn-Cunningham, B.G., Best, V., 2008. selective attention in normal and impaired hearing. Trends in Amplification 12, 283–299.
Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., Simon, J.Z., 2012. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. J. Neurophysiol. 107, 2033–2041.
Yang, X., Wang, K., Shamma, S.A., 1992. Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38, 824–839.
Yuille, A., Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. 10, 301–308.
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77, 980–991.