

Chapter 41

Robust Cortical Encoding of Slow Temporal Modulations of Speech

Nai Ding and Jonathan Z. Simon

Abstract This study investigates the neural representation of speech in complex listening environments. Subjects listened to a narrated story, masked by either another speech stream or by stationary noise. Neural recordings were made using magnetoencephalography (MEG), which can measure cortical activity synchronized to the temporal envelope of speech. When two speech streams are presented simultaneously, cortical activity is predominantly synchronized to the speech stream the listener attends to, even if the unattended, competing-speech stream is more intense (up to 8 dB). When speech is presented together with spectrally matched stationary noise, cortical activity remains precisely synchronized to the temporal envelope of speech until the noise is 9 dB more intense. Critically, the precision of the neural synchronization to speech predicts subjectively rated speech intelligibility in noise. Further analysis reveals that it is longer-latency (~100 ms) neural responses, but not shorter-latency (~50 ms) neural responses, that show selectivity to the attended speech and invariance to background noise. This indicates a processing transition, from encoding the acoustic scene to encoding the behaviorally important auditory object, in auditory cortex. In sum, it is demonstrated that neural synchronization to the speech envelope is robust to acoustic interference, whether speech or noise, and therefore provides a strong candidate for the neural basis of acoustic-background invariant speech recognition.

N. Ding
Department of Electrical and Computer Engineering,
University of Maryland, College Park, MD, USA

J.Z. Simon (✉)
Department of Electrical and Computer Engineering,
University of Maryland, College Park, MD, USA

Department of Biology, University of Maryland,
College Park, MD, USA
e-mail: jzsimon@umd.edu

1 Introduction

Normal-hearing human listeners are remarkably good at understanding speech in adverse listening environments. Acoustic degradation to speech can be due to energetic or informational masking (Brungart 2001). Here, energetic masking refers to the energetic overlap between the target speech and maskers, arising dominantly in the auditory periphery. Informational masking refers to the interference caused by the perceptual similarity between the target speech and maskers, arising dominantly in the central auditory system. In this study, energetic masking of speech is studied using a stationary noise masker with a long-term spectrum matching that of speech. The stationary noise affects the audibility of speech by severely reducing its intensity contrast, i.e., the depth of the spectro-temporal modulations (see also Stone et al. 2011). Informational masking, in contrast, is exemplified by the masking caused by a competing speech signal. In this case, both speech streams are audible and intelligible, and therefore the difficulty the listeners face is to identify which speech features are from the target speech and then to selectively process them.

We recorded neural responses from normal-hearing human listeners using MEG, which can measure cortical activity precisely synchronized to the temporal envelope of speech (Ding and Simon 2012). Based on the MEG measurements, we analyzed how the neural synchronization to speech is affected by acoustic interference, speech or noise, at different intensity levels. It is demonstrated that a robust neural representation of speech is maintained in human auditory cortex, reflecting a variety of top-down and bottom-up gain control effects in human auditory cortex.

2 The Cortical Representation of Competing Speech Streams

2.1 *Experimental Procedures*

In this competing-speech experiment (Ding and Simon 2012), two stories, narrated by a male and a female speaker respectively, were mixed into a single acoustic channel with different intensity ratios. One speaker was always presented at roughly 75 dB SPL, while the other was presented at either the same level (by RMS value) or 5 or 8 dB weaker. These two speakers were referred to as the constant-intensity speaker and the varying-intensity speaker, respectively. For this set of speech mixtures, when different speakers were attended to, the target-to-masker ratio (TMR) ranges from -8 to 8 dB. Each stimulus was presented twice under each attentional condition (attend-to-male vs. attend-to-female).

For each TMR, two 1-min duration stimuli were presented, after each of which a comprehension question was asked to ensure the subjects' attention. The listeners correctly answered 71 % of the questions, and this percentage did not significantly vary with TMR ($p > 0.7$, one-way repeated-measures ANOVA). Six subjects participated in the experiment. Five of them were asked to subjectively

rate speech intelligibility (as percentage) after the first listening to each stimulus. The neuromagnetic signals were recorded using a 157-channel whole-head MEG system.

2.2 Neural Reconstruction of Each Speech Stream

The temporal envelope of each speech stream was reconstructed separately, by integrating neural activity over time and MEG sensors. The reconstructed envelope $\hat{s}(t)$ is $\hat{s}(t) = \sum_k \sum_{\tau} r_k(t + \tau) h_k(\tau)$, where $r_k(t)$ is the recording from the k th MEG sensor and the decoder $h_k(\tau)$ is a weighting matrix for the k th MEG sensor and a time lag τ between the stimulus and response. The decoder was optimized using boosting, with ten-fold cross validation, which maximizes the accuracy of neural reconstruction, i.e., the correlation between the reconstructed envelope and the actual envelope of speech (Ding and Simon 2012). The envelope of each speaker, $s(t)$, was expressed on a linear amplitude scale. The accuracy of the neural reconstruction depends on how precisely the speech envelope is encoded in neural activity and is an index of the fidelity of neural encoding.

The chance-level reconstruction accuracy was estimated by generating pseudo-reconstructions based on unmatched stimulus-response pairs. To create a stimulus not matching the neural response, we cut the actual stimulus into eight segments, shuffled them, and concatenated the shuffled segments. A hundred pseudo-reconstructions were generated based on the shuffled stimuli, and the maximal reconstruction accuracy from the pseudo-reconstructions was used as a threshold to test the significance of normal neural reconstructions ($P < 0.01$).

Beyond the correlation analysis, a more detailed relationship between $\hat{s}(t)$ and $s(t)$ is obtained by fitting $\hat{s}(t)$ as a function of $s(t)$. This function is called the amplitude-intensity function (AIF). Since the neural reconstruction is just the spatial-temporally integrated neural response, the AIF describes the relationship between the instantaneous amplitude of the neural response and the instantaneous intensity of the stimulus.

2.3 Results

The temporal envelope of each speaker in the stimulus is reconstructed separately from the cortical response. The correlation between the reconstruction and the actual envelope of speech is shown in Fig. 41.1a. The reconstruction accuracy is above chance for both speech signals ($P < 0.01$ for every condition) and is significantly higher for the attended speech ($P < 0.02$, 2-way repeated-measures ANOVA, factors: attention, TMR). The main effect of TMR and the interaction were both not significant. Since the same decoder is used in every TMR condition, the TMR-independent reconstruction accuracy suggests TMR-independent neural encoding of the temporal modulations of each speech stream.

The TMR-independent neural reconstruction implies neural compensation for the intensity change of the speakers. This is further investigated using the AIF, which describes the relationship between the instantaneous amplitude of the neural response and the instantaneous intensity of the stimulus envelope (Fig. 41.1b). The AIFs for the two speakers show distinct behaviors. The AIF for the varying-intensity speaker shifts leftwards as the intensity of the speaker decreases, regardless of the attentional state of the listener. A leftward shift of the AIF indicates an increase in response gain since lower intensity is needed to achieve a given response amplitude. When fitted by a line, the AIF shifts 6.0 ± 0.2 dB and 5.0 ± 1.1 dB (Mean \pm SEM) for the attended and unattended speaker, respectively, as the intensity of the speaker changes by 8 dB. The AIF for the constant-intensity speaker, in contrast, is not significantly affected by the intensity change of varying-intensity speaker. Therefore, the neural representation of each speaker only adapts to the mean intensity of that speaker, rather than the mean intensity of the stimulus mixture. In other words, neural adaptation to sound intensity is auditory stream specific.

3 Cortical Representation of Speech Masked by Noise

3.1 Methods

In the speech-in-noise experiment (Ding and Simon 2013), each stimulus consisted of a 50-s duration spoken narrative. Stationary noise matching the long-term spectrum of speech was generated using a 12th-order linear predictive model and mixed into speech with one of the following six TMRs: quiet, +6, +2, -3, -6, and -9 dB. The intensity of speech was the same for all stimuli while the intensity of the noise varied. Ten subjects participated.

Each stimulus (12 in total) was presented three times. The TMR always increased or decreased every two sections (counterbalanced over subjects). The subjects were asked a comprehension question after each section. During the first presentation of each stimulus, the subjects were asked to rate the intelligibility of each stimulus. The order how the sections were presented, whether with increasing or decreasing TMR, did not affect speech intelligibility (two-way repeated-measures ANOVA, factors: TMR, Order) or the neural reconstruction of speech (the same ANOVA) and therefore was not distinguished in the analysis.

3.2 Results

To investigate how the cortical representation of speech is affected by background noise, we reconstructed the temporal envelope of the underlying clean speech, not

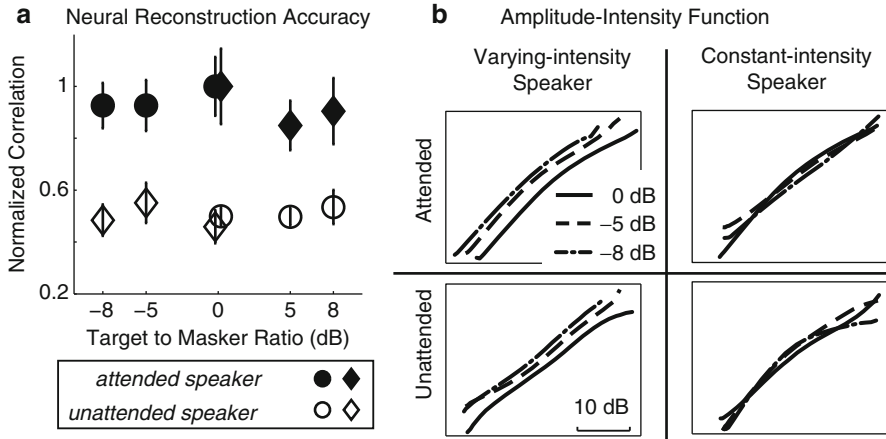


Fig. 41.1 Cortical reconstruction of each of two competing speech streams. **(a)** The correlation between neural reconstruction and the actual envelope of each speaker (*filled* and *hollow* symbols for attended and unattended speakers, respectively). The two speakers are shown by *circles* and *diamonds*, respectively. The correlation with each speaker is normalized based on the correlation at 0 dB TMR when the speaker is attended. **(b)** The AIF for each speaker, under each attentional state. The *x*-axis is the stimulus intensity and the *y*-axis is the dimensionless amplitude of the neural reconstruction on a linear scale. The level difference between the speakers is indicated by the *line* style

the actual stimulus, from the cortical response to a noisy stimulus (Fig. 41.2). The correlation between the neural reconstruction and the actual envelope of speech remains high until the TMR drops to -9 dB (Fig. 41.2). This indicates that, above -9 dB, the temporal modulations of speech are cortically encoded by phase-locked activity, regardless of the degradation caused by noise. Decoding accuracy was not affected by TMR when the -9 dB condition is excluded (2-way repeated-measures ANOVA, factors: TMR, Trial).

At the intermediately low TMR of -3 dB, the median of the rated speech intelligibility was 55 % and varied widely. At this TMR, individual subject's subjectively rated speech intelligibility is significantly correlated with neural reconstruction accuracy ($R=0.78\pm 0.15$, bootstrap, Fig. 41.2). No such correlation was found at high and low TMRs, because of ceiling (median >90 %) and floor (median ≤ 10 %) effects in the ratings.

Stationary background noise reduces the depth of the spectro-temporal modulations, i.e., intensity contrast, of speech. Therefore, the robust neural encoding of speech suggests that the loss of stimulus contrast is compensated for by the auditory system. To demonstrate this, we estimated the AIF for each TMR condition and found the AIF to be strongly TMR dependent (Fig. 41.2), showing neural adaptation to intensity contrast. The slope of the AIF, extracted by a linear regression, increases 16 ± 2 dB (Mean \pm SEM) as TMR decreases from infinity (quiet) to -6 dB.

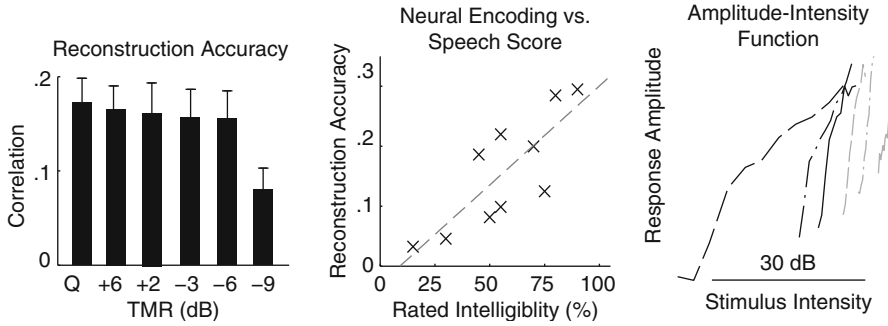


Fig. 41.2 Neural reconstruction of speech masked by stationary noise. (*Left*) The correlation between the neural reconstruction and the temporal envelope of the underlying clean speech. (*Middle*) At -3 dB TMR, individual subject's intelligibility rating is significantly correlated with the accuracy of neural reconstruction. (*Right*) The AIF for each TMR condition. The curves, from left to right, correspond to conditions with decreasing TMR

4 Time Course of Neural Encoding During Energetic and Informational Masking

4.1 Temporal Response Function

How the cortical response is generated by temporal modulations of speech can be modeled using the temporal response function (TRF). For the competing-speech experiment, the response was modeled as the sum of two TRF models, one for each speaker. The TRF was derived from a spectro-temporal response function that was estimated based on the normalized spectrogram of speech (z-score) expressed on a logarithmic scale (Ding and Simon 2012). Since the TRF is based on the normalized spectrogram, the gain of the TRF is invariant to the changes of the stimulus intensity only if neural activity is fully adapted to such changes. A TRF is estimated for each MEG sensor. Two major components are seen in the TRF, with distinct neural sources and response latencies. As an effective representation of both components, the TRF shown here is the sum of the TRFs projected to the neural sources of the two components.

4.2 Time-Dependent Gain Control

The TRFs from the competing-speech and speech-in-noise experiments are shown in Fig. 41.3. The TRF characterizes the temporal evolution of the neural response evoked by a unit power increase of the speech stream it is applied to.

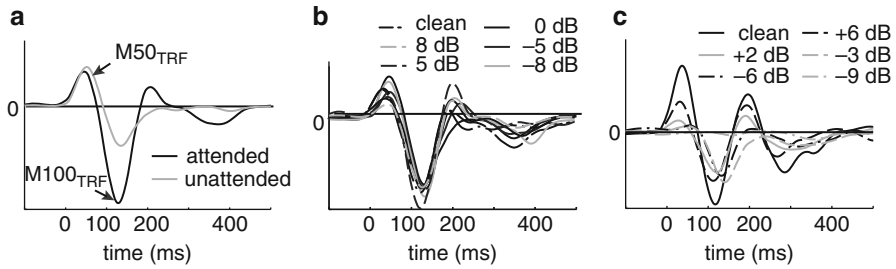


Fig. 41.3 The temporal response function (*TRF*) in the competing-speech experiment (**a–b**) and in speech-in-noise experiment (**c**). (**a**) The *TRF* for the attended and unattended speech streams. The $M100_{TRF}$ is modulated by attention, but the $M50_{TRF}$ is not. (**b**) The shape of the *TRF* is invariant to TMR. (**c**) In noise, the $M50_{TRF}$ weakens as the TMR decreases while the amplitude of $M100_{TRF}$ remains stable as the TMR changes from +6 to –6 dB

The earliest two peaks of the *TRF* have latencies near 50 and 100 ms, called the $M50_{TRF}$ and $M100_{TRF}$, respectively. In the competing-speech experiment, the amplitude of the $M100_{TRF}$ but not the $M50_{TRF}$ is significantly modulated by attention ($P < 0.03$, four-way repeated-measures ANOVA, factors: attention, hemisphere, speaker, and TMR). Furthermore, the shape of the *TRF* is independent of the TMR, for both the constant-intensity speaker and the varying-intensity speaker. A stimulus-invariant *TRF* reflects a complete neural adaptation to the mean intensity of the stimulus, since the *TRF* is derived from normalized speech envelope. In the speech-in-noise experiment, the $M50_{TRF}$ weakens as TMR decreases (negative correlation between $M50_{TRF}$ amplitude and TMR, $P < 0.001$, bootstrap). The $M100_{TRF}$ however, remains largely stable between +6 dB TMR and –6 dB TMR (no significant correlation between $M100_{TRF}$ amplitude and TMR).

5 Spectro-temporal Processing of Speech

5.1 Object-Based Gain Control

Neural adaptation to sound intensity occurs at multiple stages of the auditory system (Robinson and McAlpine 2009). The competing-speech experiment further suggests that the neural adaptation to sound intensity occurs separately for each auditory stream/object. In that experiment, the stable representation of the varying-intensity speaker must be maintained by neural adaptation to sound intensity, while the stable representation of the constant-intensity speaker requires no adaptation to the overall intensity of the sound mixture (which itself covaries with the intensity of the varying-intensity speaker). Therefore, the stable representation of

both speakers cannot be explained by a simple mechanism of global intensity gain control, which would result in the neural representation of both speakers to be modulated in the same way based on the overall intensity of the acoustic stimulus. Instead, the results suggest object-specific intensity gain control.

5.2 Latency of Gain Control Effects During Energetic and Informational Masking

The properties of the acoustic masker, i.e., whether informational or energetic, influence the cortical processing of speech differentially. As revealed by the TRF analysis, in the competing-speech experiment, both the shorter- and longer-latency cortical responses are insensitive to the change in masker intensity. In the speech-in-noise experiment, however, only the longer-latency response $M100_{\text{TRF}}$ is resilient to the masker. This influence of masker property is straightforward to explain. During informational masking, the audibility of each stream of speech is seldom a problem and therefore each stream drives early auditory response effectively. The key question during informational masking is the selection of acoustic features belonging to the speech target, which is only reflected by the attentional modulation of long-latency responses. During energetic masking, however, background noise reduces the audibility of speech and therefore attenuates the shorter-latency response. The noise robustness of the long-latency response can only be maintained by additional active neural processing. In summary, the shorter-latency (~ 50 ms) response from core auditory cortex mainly reflects the audibility of a sound stream, while the longer-latency (~ 100 ms) response is a robust representation of the target speech stream.

5.3 Relation to Speech Intelligibility

On the one hand, the cortical synchronization to speech is more robust to acoustic interference than rated speech intelligibility. For example, at -6 dB TMR, cortical synchronization to speech is not affected by acoustic interference, whether speech or noise, but rated speech intelligibility drops to about 50 % for a speech masker and only about 10 % for a noise masker. On the other hand, the precision of cortical synchronization to speech is a good predictor of individual's intelligibility rating, when speech is masked by noise at -3 dB TMR. We explain this difference by dividing speech recognition into two consecutive processes. One is the parsing of the continuous and possibly noisy acoustic input into basic processing unit, e.g., syllables. The other is the decoding of linguistic information from each unit. We argue that the MEG-measured cortical synchronization to speech reflects the first parsing process, which is more reliable than the decoding of phonemic information from

each unit (see Woodfield and Akeroyd 2010, for psychoacoustical evidence). In the presence of an intermediate amount of noise, the parsing process becomes a bottleneck for speech recognition, and therefore listeners who are better at extracting basic speech units rate speech intelligibility as higher.

6 Conclusion

We found that large-scale coherent cortical activity is precisely synchronized to the temporal modulations of speech, even in the presence of an acoustic masker twice as intense as the speech target. Two major sources of the speech-synchronized neural response are identified. One has shorter latency (~50 ms) and is from roughly core auditory cortex. The other has longer latency (~100 ms) and is from posterior association auditory cortex. The shorter-latency response is not modulated by attention and is susceptible to background noise. The longer-latency response, however, is strongly modulated by attention and is resilient against acoustic interference. In summary, the results suggest the emergence of a neural representation of the target speech stream embedded in a complex auditory scene. This auditory stream-specific representation is enhanced, from shorter-latency to longer-latency neural responses and from core to posterior auditory cortex.

Acknowledgments We thank NIH grant R01 DC 008342 for support.

References

- Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109:1101–1109
- Ding N, Simon JZ (2012) The emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci USA* 109(29):11854–11859
- Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background insensitive cortical representation of speech. *J Neurosci* 33:5728–5735
- Robinson BL, McAlpine D (2009) Gain control mechanisms in the auditory pathway. *Curr Opin Neurobiol* 19:402–407
- Stone MA, Fullgrabe C, Mackinnon RC, Moore BCJ (2011) The importance for speech intelligibility of random fluctuations in “steady” background noise. *J Acoust Soc Am* 130:2874–2881
- 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