Noise Robust Neural Synchronization to Speech Envelope

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

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

Computational Sensorimotor Systems Lab

Introduction

The neural underpinnings of noise-robust perception of speech

Cortical activity is precisely synchronized to the envelope of speech, and we hypothesize that this cortical synchronization is robust to noise, and provides a neural basis for noise-robust perception of speech. We tested this hypothesis by recording from human subjects listening to a narrated story in noise, and demonstrated that the low frequency cortical synchronization to speech is indeed robust to noise over a broad signal-to-noise ratio (SNR) range.

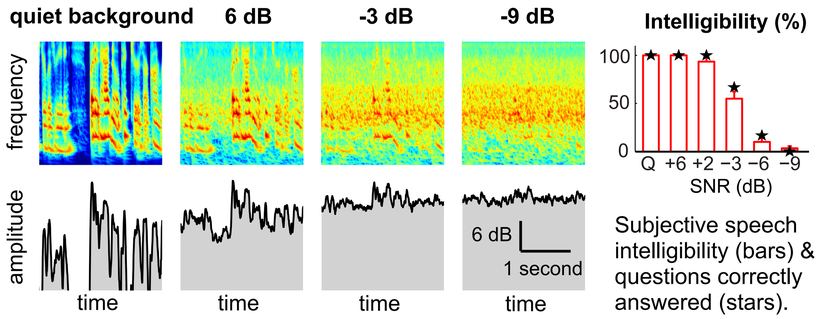
Stimuli & Data Analysis

Stimuli

The speech material was selected from a narrated story, each section 50 seconds in duration. Spectrally matched stationary noise was mixed into speech with one of six SNRs: quiet (no noise added in), +6 dB, +2 dB, -3 dB, -6 dB, and -9 dB.

All sections were presented sequentially and then repeated twice (3 trials in total). The subjects answered a comprehension question after each presentation, and rated the speech intelligibility after the first presentation of each section.

The background noise reduces the dynamic range of the stimulus (as evident from the stimulus envelope), and distorts the spectro-temporal features of speech.

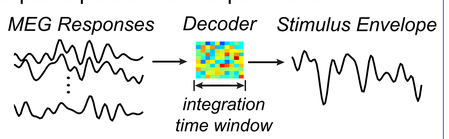


Subjective speech intelligibility (bars) & questions correctly answered (stars).

Data Analysis

MEG: 157-channel, whole-head MEG; 1 kHz sampling rate, resampled to 40 Hz. The neural source of MEG activity is localized using a equivalent current dipole model, one per hemisphere. 10 subjects participated in the experiment.

Neural Reconstruction: We reconstructed the envelope of speech using a linear decoder that integrates MEG activity over time and sensors.



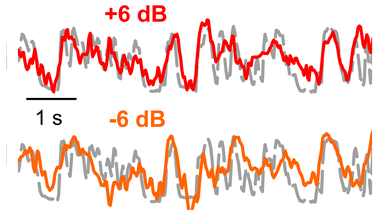
TRF: The neural response in each MEG sensor is modeled by the stimulus envelope convolved with a temporal response function (TRF). The TRF models the neural response evoked by a unit power increase of the stimulus. Both the decoder in the reconstruction analysis and the TRF were estimated using boosting with 10-fold cross validation.

Conclusions

- Even in the presence of stationary noise, auditory cortical activity is reliably synchronized to the envelope of speech, suggesting a noise-robust neural representation of the syllabic/phrasal rhythm of speech.
- A more robust neural representation is observed for lower (e.g. 2 Hz) rather than higher (e.g. 6 Hz) frequency responses, for more posterior auditory cortical areas rather than closer to core auditory cortex, and for longer (e.g.100 ms) rather than shorter (e.g. 50 ms) latencies.
- Such cortical synchronization predicts individual's speech score in noise, and is a possible neural basis for noise-robust speech recognition.

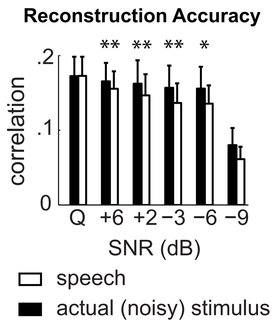
Neural Reconstruction of Speech Envelope

The envelope of speech, not the noisy stimulus, is reconstructed from the neural response to each noisy stimulus. Examples of the reconstruction (red & orange) are shown on the right, together with the true envelope of speech (dashed gray). The reconstruction is similarly accurate at +6 dB and -6 dB.



The neural reconstruction accuracy is evaluated by the correlation between the reconstruction and actual speech envelope (black bars on the left). The reconstruction accuracy is not significantly affected by SNR above -6 dB SNR.

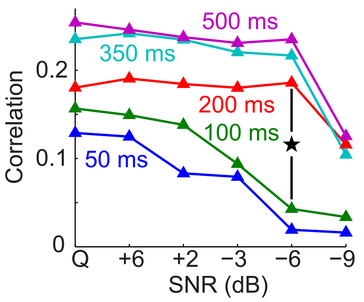
A separate reconstruction analysis is applied to reconstruct the envelope of the actual noisy stimulus (white bars). This reconstruction, although straightforward, is not as accurate as the reconstruction of speech envelope, indicating a neural encoding of the embedded speech rather than the actual noisy stimulus.



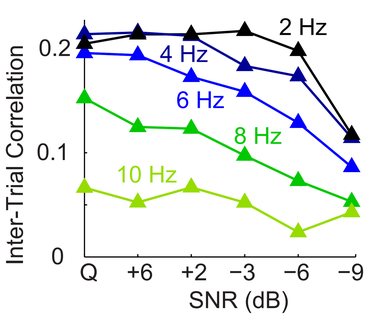
Cortical activity is precisely synchronized to the envelope of speech and this synchronization is not degraded by noise until -9 dB SNR.

Long-term Temporal Integration

Reconstruction Accuracy by Window Integration Time



Neural Phase-locking by Response Frequency



In the reconstruction analysis, the speech envelope at a given time moment is reconstructed by integrating neural activity over a time window. The reconstruction accuracy improves as the window size increases, and saturates when the window is 500 ms in duration.

The largest increase of reconstruction accuracy is seen at -6 dB, when the window size increases from 100 ms to 200 ms. The reconstruction accuracy is SNR-independent (above -6 dB), only if the window size is longer than 100 ms.

Similarly, only very low frequency, e.g. 2 Hz, neural activity is synchronized to speech in a SNR-independent manner (above -6 dB), while higher frequency, e.g. 6 Hz, neural synchronization is more susceptible to noise.

Noise robust neural synchronization to speech requires long-term integration, and is only reflected by very low frequency neural activity.

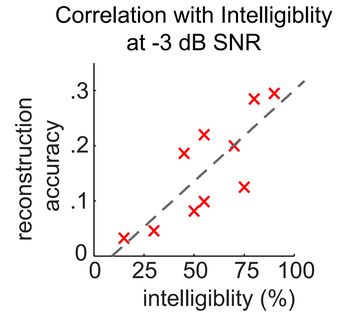
References: Details of the analysis can be found in Ding & Simon (J Neurophys. 2012) and Ding & Simon (PNAS, 2012).

Acknowledgement: work supported by NIH R01 DC-008342.

Relation between Neural Synchronization and Speech Intelligibility

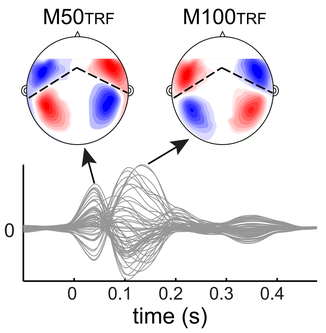
At -3 dB SNR, the median speech intelligibility is about 50%. At this SNR, the accuracy of neural synchronization is significantly correlated with the subjectively rated intelligibility cross subjects (r = 0.79). (At other SNRs the median speech intelligibility is > 90% or ≤10%.)

The accuracy of neural synchronization also predicts the SRT of individuals (the SNR when intelligibility drops to 50%, r = 0.63).



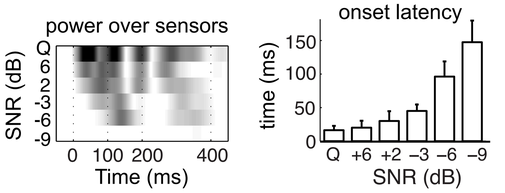
The accuracy of neural synchronization predicts individual speech score in noise (-3 dB SNR).

Temporal Response Function



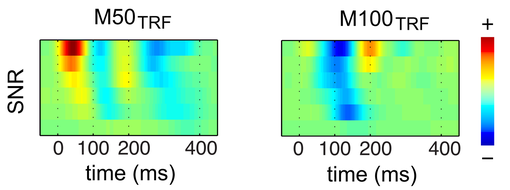
A temporal response function (TRF) is estimated for each MEG sensor. It represents the neural response evoked by a unit power increase of the stimulus. The TRF has two salient peaks, the M50TRF and M100TRF, which have opposite polarity. The source of the M100TRF is consistent with the source of the M100 evoked by a tone pip, which is in posterior association auditory cortex. The source of the M50TRF is more anterior than the source of the M100TRF, and is more close to core auditory cortex.

Power of the TRF over all MEG Sensors



The onset latency of the TRF is delayed as the SNR decreases. The amplitude, however, is relatively stable between -6 and 6 dB SNR.

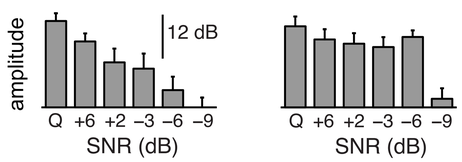
TRF at the Neural Source Locations of the M50TRF & M50TRF



Projections of the TRF to the source locations of the M50TRF and M100TRF

The M100TRF polarity is consistent with that of of the M100 and is defined to be negative.

Amplitude of the M50TRF & M50TRF



The amplitude of the M50TRF (left) continuously decreases with SNR while the amplitude of the M100TRF (right) is stable above -9 dB.

Longer-latency (~100 ms) responses from posterior auditory cortex is robust to noise, but not shorter-latency (~50 ms) responses from roughly core auditory cortex.