**Neural Entrainment to Speech Envelope in response to Perceived Sound Quality**

Dat Quoc Ngo, Garret Oliver, Gleb Tcheslavski, and Chin-Tuan Tan, Member, IEEE[[1]](#footnote-1)

***Abstract –* The extent, to which people listen to and perceive the speech content at different noise levels varies from individual to individual. In past research projects, the speech intelligibility was determined by rating assessment, which suffered from variation of subjects’ physical features. The purpose of this study is to investigate electroencephalography (EEG) by implementing multi-variate Temporal Response Function (mTRF) to examine neural responses to speech stimuli at different sound and noise levels. The result of this study shows that the front-central area of the brain clearly shows the envelope entrainment to speech stimuli.**

# Introduction

Speech perception in our auditory system is a hierarchical system composed of multiple stages that is consistently responding to speech tokens (e.g. phoneme, words and sentences) regardless of their acoustic variations caused by internal (e.g. hearing impairment and foreign accents) or external (e.g. background noise) sources of interferences. The challenge for such perception system to work in noisy environment has always been the perceived quality of the acoustic speech signal and the background noise itself. A direct consequence in identifying context and extracting meaning from a degraded acoustic speech signal is an increased cognitive processing and associated listening effort.

In order to cope with the speed and efficiency required in everyday conversation in noisy environments, listeners are forced to put in a greater listening effort in relying to a greater extent of cognitive processes to successfully comprehend an acoustically degraded speech, when the perceived sound quality does not match up the listener’s expectation that is dependent on the quality of acoustic speech signal. These processes, often referred to as “domain general”, are often associated with executive tasks that occur similarly regardless of the input modality of a stimulus, for instance, the 3-step sound quality assessment process [2], which consists of perception, judgement, and description. Recent evidence had shown that auditory processing relies on domain-general cognitive processes to a greater degree when speech input is acoustically degraded and reflects as changes in the associated neural processing. A previous study by Di Liberto and colleagues [7] had shown that neural activities in response to continuous speech captured using EEGs that correlated to the spectro-temporal information and various speech tokens in the continuous speech.

In this study, we attempted to examine the mapping between the spatiotemporal information of continuous speech and its corresponding neural responses, occurring usually at longer latency due to the hierarchical processing, when listening in an adverse environment, and its relationship to the perceived quality of speech. In doing so, we sought to determine if EEG reflects a neural activity that is passively following the acoustic energy of speech (usually represented by speech envelope) under a noisy environment when the perceived sound quality of the speech is poor.

# Experiment

To investigate the relationship between neural activities and sound stimuli, we fitted mTRF’s [1] between the speech stimuli and the corresponding neural responses [4] measured using EEGs [3] via linear regression. The mTRF model actually captures the cortical activities that covary with changes in stimuli, which is the fluctuation of speech envelope in this study. Alternatively, the mTRF model could also be implemented in the opposite or backward direction to generate models that offer a way to investigate how speech stimuli are encoded in neural responses.

# *Participants*

This study was approved by the Lamar University IRB committee. Eight native English speakers (5 males and 3 females) with normal hearing (verified at a local hearing clinic) of age from 20 to 24 have participated in this study. All participants reported no history of neurophysiological disorders and were not tired or sleepy at the time of data acquisition.

### *Stimuli*

Four isolated speech sentences spoken by a male speaker and other isolated speech sentences spoken by a female speaker were randomly excerpted from two separate audio books. It was unlikely that stimuli provoked the emotional response and to cause changes in EEG production among participants. The audio stimuli were adjusted to three sound pressure levels (SPL’s): (75, 65 and 55 dB). Two types of noise, White and Babble noise, were generated at three SPL’s (75, 65 and 55 dB) and mixed with the speech to produce stimuli at different signal-to-noise ratios (SNRs) and speech levels. Note that speech at 75dB SPL and noise at 65dB SPL was considered a different listening condition than a speech at 65dB SPL and noise at 55 dB SPL, even though both stimuli had same of 10dB. Therefore, the stimuli pool consisted of 24 original and 96 noisy speech fragments.

### *Perceived sound quality rating*

Each participant was instructed to rate the perceived sound quality of stimuli that were presented in a randomized order. The listeners were ranking the apparent quality of speech using the scale from 1 (inferior) to 10 (excellent).

### *EEG recording and Pre-processing*

Audible stimulation was delivered diotically to the participants via Etymotic insert earphones (Etymotic Research ER3A, 10Ω impedance), to evoke the corresponding cortical response which is captured by the EEG recording. Each EEG recording was processed in synchrony with the stimulus.

EEG was recorded via the ASA-Lab40 acquisition system by ANT Neuro, Netherlands. Continuous EEG was pre-filtered in 0.3-50 Hz range, notch-filtered at 60 Hz, sampled at 1,024 Hz, and recorded from 32 electrodes positioned according to the extended International 10/20 placement map. EEG was processed offline by first fragmenting it into epochs synchronized with stimulation. Each epoch was baseline-corrected and filtered with a CAR spatial filter to reduce surface currents.

##### *Analysis and Synthesis with mTRF Models*

Regularized Linear Regression algorithm was used to train mTRF models [1]. Unlike univariate Temporal Response Function models implementing single features (e.g. electrodes), mTRF models that relate EEG responses at different electrodes of participants to the corresponding speech envelopes in linear time lags (moments). The models reflect the multi-feature hierarchical sound processing. In the forward direction, the model and original speech envelopes can be used to predict EEG responses. Similarly, in the backward direction, the model, and original EEGs can be used to predict original envelopes of sentences.

For every subject, the EEG data and speech envelopes were predicted by mTRF models [1] and were tested with the subject’s original EEG data and speech envelopes. We employ this self-validation approach to quantify changes of EEG responses against speech stimuli. The quality of the prediction was assessed using correlation (RHO) and mean squared error (MSE) [1].

# Result

The results of the study are illustrated using one fragment spoken by a male speaker at 75 and 65 dB SPL and the same fragment mixed with a babble noise at 55 and 75 dB SPL to produce two noisy speech stimuli at SNR of +20 and -10 dB SNR. The corresponding EEG data (neural response) of the one participant was used to train four separate stimulus-response models (mTRFs) for these four fragments. The stimulus-response models were then used to reconstruct EEGs from sound stimuli. The reconstructed neural signals were compared with the real EEG data of the listener. Likewise, the same analysis was performed with the rest of 120 sentences. For forward modeling, the stimulus-response model with TRF’s in the form of matrix, linearly maps the stimulus and the neural response. Whereas for backward modeling, a reverse stimulus-response mapping was derived. For simplicity, only the results of the fifth participant (subject 5) was described in figures 1 and 2.

##### *EEG Prediction*

Previous study by Ding and Simon [5] with similar experimental setup had shown that predicted EEG can be reconstructed from the mathematically mapped EEG-stimulus STRF (spectro-temporal response function) models with low errors. Instead STRF, mTRF (multi-variate response function) was applied to quantify the EEG responses of different channels against speech stimuli at various frequency bands [1]. The EEG data predicted from mTRF models closely followed the average trend of the original EEG morphology, as illustrated in Fig. 1 and 2, where the predicted EEG indicated by the red lines and original EEG indicated by green lines. The predicted EEG was closely following the energy fluctuation of the original EEG, regardless of level of babble noise present.

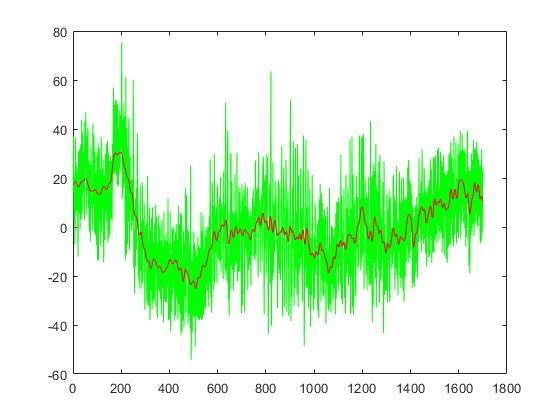
The accuracy of the predicted EEG in following the original EEG was measured by computing the mean square error (MSE) between these two EEG morphologies, as tabulated in Table 1 and 2. In general, high accuracies were observed across all sentences and subjects. Even with the noisy sentence at -10dB SNR, where speech at 65dB SPL and babble noise at 75dB SPL), the MSE across subjects were ranging between 1.07×10-3 and 2.40×10-3.

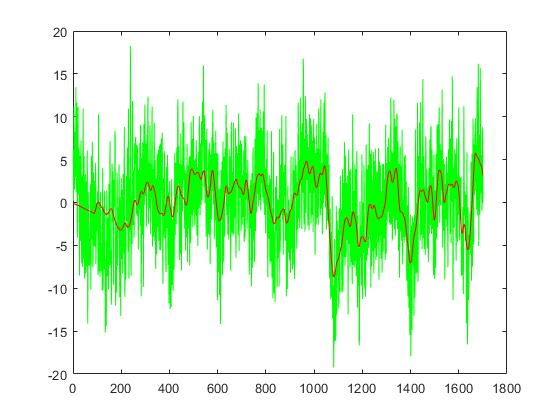
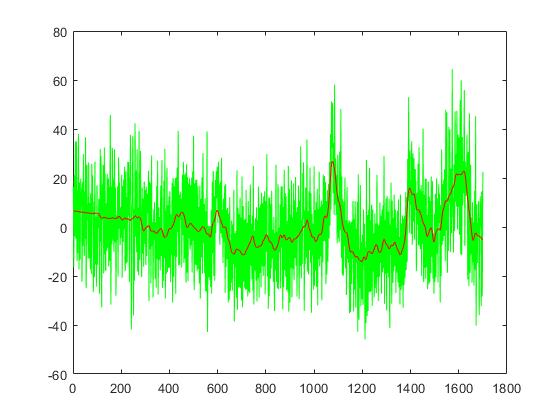
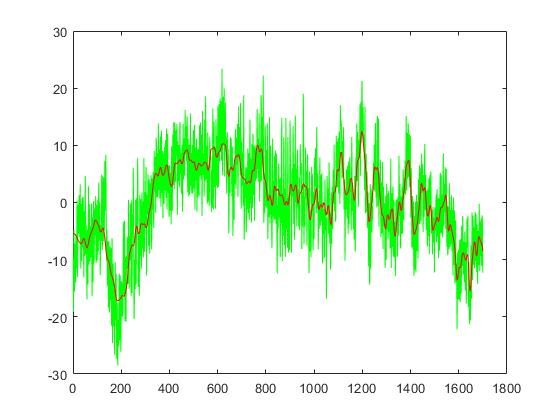
##### *Front-central Channels*

With the four illustrating sentences described at the beginning of the result section, the predicted EEG and original EEG elicited by the four sentences were compared in four separate plots at Fz and Cz positions, which located at the front central area of the scalp, in Fig. 1 and 2 respectively. Correspondingly, the MSEs between predicted EEGs and original EEGs in Table 1 and 2. Looking at the subject 5 of Table 1, the MSEs computed at Cz position were 27.9×10-3 with original sentence spoken at 65dB SPL and 21.6×10-3 with noisy sentence at -10 dB SNR (speech at 65dB SPL and babble noise at 75dB SPL). Similarly, in Table 2, the corresponding MSE’s computed at Fz position were 313×10-3 and 209×10-3, respectively.

# Discussion

We found that cortical entrainment to speech envelope was better observed at front-central electrodes. In our





(A) (B)

(C) (D)

Figure 2. Predicted EEGs (red) and original EEGs (green) for subject 5 listening to a sentence spoken by a male-speaker (a) at 65dB SPL at Fz position; (b) at 65dB SPL with babble noise at 75dB SPL at Fz position; (c) at 65dB SPL at Cz position; (d) at 65dB SPL with babble noise at 75dB SPL at Cz position.

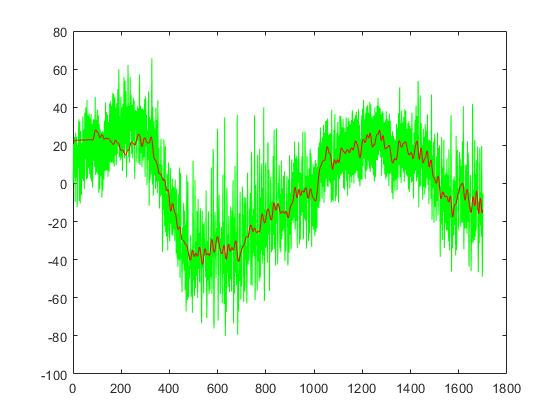
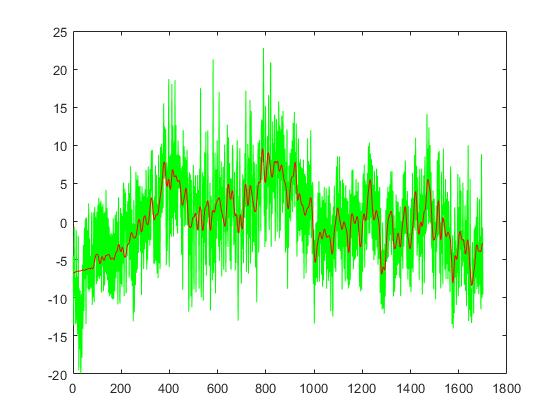
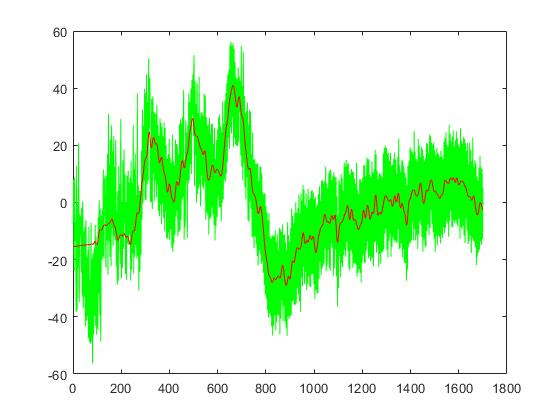
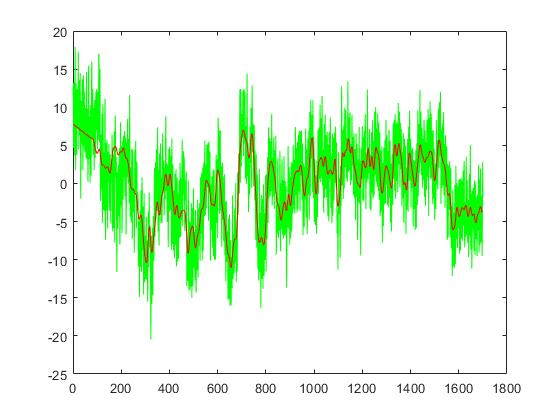


Figure 1. Predicted EEGs (red) and original EEGs (green) for subject 5 listening to a sentence spoken by a male-speaker (a) at 75dBSPL at Fz position; (b) at 75dB SPL with babble noise at 55dB SPL at Fz position; (c) at 75dB SPL at Cz position; (d) at 75dB SPL with babble noise at 55dB SPL at Cz position.

a.u.

a.u.

a.u.

a.u.

a.u.

a.u.

a.u.

a.u.

Times (ms) Time (ms)

Times (ms) Time (ms)

Times (ms) Time (ms)

Times (ms) Time (ms)

(A) (B)

(C) (D)

Predicted EEG

Read EEG

Predicted EEG

Read EEG

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Speech at 75dB SPL | | | | | | | | |
| MSE (×1E-3) | 171 | 413.72 | 484.04 | 565.76 | 160.93 | 270.04 | 207.66 | 344.9 |
| Sound Quality | 10 | 10 | 6 | 6 | 10 | 10 | 10 | 7 |
| Speech at 75dB SPL + Babble noise at 55dB SPL | | | | | | | | |
| MSE (×1E-3) | 160.02 | 400.57 | 508.9 | 563.46 | 232.41 | 272.65 | 213.80 | 341.1 |
| Sound Quality | 9 | 9 | 7 | 10 | 6 | 8 | 9 | 6 |
| Speech at 65dB SPL | | | | | | | | |
| MSE (×1E-3) | 184.39 | 488.74 | 499.24 | 589.07 | 313.11 | 275.21 | 246.43 | 262.2 |
| Sound Quality | 10 | 8 | 5 | 10 | 10 | 10 | 10 | 7 |
| Speech at 65dB SPL + Babble noise at 75dB SPL | | | | | | | | |
| MSE (×1E-3) | 161.45 | 404.91 | 491.09 | 556.91 | 208.95 | 271.08 | 210.32 | 345.5 |
| Sound Quality | 3 | 2 | 1 | 2 | 1 | 3 | 3 | 2 |

TABLE 1. Accuracy and sound quality of EEG prediction across participants for sentences at 65 & 75dB SPL and same sentences with babble noise at 55 & 75dB SPL at EEG location (Cz) from male speakers

TABLE 2. Accuracy and sound quality of EEG prediction across participants for sentences at 65 & 75dB SPL and same sentences with babble noise at 55 & 75dB SPL at EEG location (Fz) from make speakers

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Speech at 75dB SPL | | | | | | | | |
| MSE (×1E-3) | 10.02 | 11.26 | 34.69 | 53.33 | 15.09 | 7.43 | 60.69 | 168.6 |
| Sound Quality | 10 | 10 | 6 | 6 | 10 | 10 | 10 | 7 |
| Speech at 75dB SPL + Babble noise at 55dB SPL | | | | | | | | |
| MSE (×1E-3) | 9.84 | 11.81 | 33.98 | 56.03 | 20.35 | 10.51 | 91.28 | 233.3 |
| Sound Quality | 9 | 9 | 7 | 10 | 6 | 8 | 9 | 6 |
| Speech at 65dB SPL | | | | | | | | |
| MSE (×1E-3) | 40.19 | 15.66 | 39.85 | 60.86 | 27.89 | 12.22 | 113.06 | 262.2 |
| Sound Quality | 10 | 8 | 5 | 10 | 10 | 10 | 10 | 7 |
| Speech at 65dB SPL + Babble noise at 75dB SPL | | | | | | | | |
| MSE (×1E-3) | 10.74 | 11.67 | 30.66 | 52.44 | 21.61 | 9.37 | 82.38 | 239.7 |
| Sound Quality | 3 | 2 | 1 | 2 | 1 | 3 | 3 | 2 |

experiment, the MSE’s between the predicted EEG and original EEG computed at front-center area of the scalp were smaller than those computed at other parts of the scalp. This observation was consistent with the outcome reported by a previous study (Ding and Simon, 2014), where they reported the intensity and number of neural activities recorded at FCz, one of front-central channels were greater than those collected at non-front-central channels. For our findings, the outperforming prediciton at front-central electrodes and the similarity of the predicted and original waveforms denote that the information of the envelope entrainment to continuous speech could be mostly stored at front-central head scalp.

In the follow-up study, we will continue to work towards achieving a more in-depth understanding about neural activities that entrain to speech, and hopefully, inspiring novel technique to predict EEGs from speech, or vice versa, for a possible optimization of signal processing in our hearing devices supervised by brain activities.

References

[1] M. Crosse, G. Liberto, A. Bednar, E. Lalor, “The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli,” *Frontiers in Human Neuroscience*, vol. 10, 2016. [Online serial]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2016.00604/full. [Accessed Oct. 20, 2018].

[2] U. Jekosch, *Voice and speech quality perception: assessment and evaluation*. 2003. [E-book] Available: Springer Science & Business Media.[E [3] R. Oostenveld, and P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,”

*Clinical Neurophysiology*, vol. 112, no. 4, pp. 713-719, 2001. [Online serial]. Available: https://www.clinph-journal.com/article/S1388-2457(00)00527-7/fulltext. [Accessed Oct. 20, 2018].

[4] V. Marmarelis, *Nonlinear Dynamic Modeling of Physiological Systems*. Reading, Wiley-IEEE Press, 2004. [E-book] Available: Wiley Online Library.

[5] N. Ding, and J. Simon, “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening”. *Journal of Neurophysiology,* vol.107, pp. 78-89, 2012. [Abstract]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21975452. [Accessed Oct. 5, 2018].

[6] N. Ding, and J. Simon, “Cortical entrainment to continuous speech: functional roles and interpretations.” *Frontiers in Human Neuroscience*, vol. 8, May, 2014. [Online serial]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00311/full. [Accessed Oct. 5, 2018].

[7] G. Liberto, J. O’Sullivan, E. Lalor, “Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing.” *Current Biology*, vol. 25, pp. 2457-2465. [Online serial]. Available: https://www.cell.com/current-biology/fulltext/S0960-9822(15)01001-5?\_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0960982215010015%3Fshowall%3Dtrue. [Accessed Oct. 15, 2018].

1. \* Research supported by the University of Texas at Dallas

   Dat Quoc Ngo is with the University of Texas at Dallas, Richardson, TX

   75080 USA (corresponding author to provide email: dqn170000@utdallas.edu). [↑](#footnote-ref-1)