Contribution of RMS-level-based Speech Segments to Target Speech Decoding under Noisy Conditions

Lei Wang [1,2], Ed X. Wu [2], Fei Chen [1]

[1] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

[2] Department of Electrical and Electronic Engineering, The University of Hong Kong, HK SAR, China

fchen@sustech.edu.cn

Abstract

Human listeners can recognize target speech streams in complex auditory scenes. The cortical activities can robustly track the amplitude fluctuations of target speech with auditory attentional modulation under a range of signal-to-masker ratios (SMRs). The root-mean-square (RMS) level of the speech signal is a crucial acoustic cue for target speech perception. However, in most studies, the neural-tracking activities were analyzed with the intact speech temporal envelopes, ignoring the characteristic decoding features in different RMS-level-specific speech segments. This study aimed to explore the contributions of high- and middle-RMS-level segments to target speech decoding in noisy conditions based on electroencephalogram (EEG) signals. The target stimulus was mixed with a competing speaker at five SMRs (i.e., 6, 3, 0, -3, and -6 dB), and then the temporal response function (TRF) was used to analyze the relationship between neural responses and high/middle-RMS-level segments. Experimental results showed that target and ignored speech streams had significantly different TRF responses under conditions with the high- or middle-RMS-level segments. Besides, the high- and middle-RMS-level segments elicited different TRF responses in morphological distributions. These results suggested that distinct models could be used in different RMS-level-specific speech segments to better decode target speech with corresponding EEG signals.

Index Terms: RMS-level speech segments, speech decoding, signal-to-masker ratios, EEG

1. Introduction

Human listeners have the ability to easily attend to the target speech stream in a complex auditory scene, and Cherry's study first described this ability of target speech identification in behavioral tests [1]. Recent studies have focused on the underlying neural mechanisms of the cocktail party problem with many electrophysiological measurements, such as the electrocorticogram (ECOG), magnetoencephalogram (MEG), and electroencephalogram (EEG) [e.g., 2-4]. Compared with the ECOG and MEG methods, EEG signals could be easily obtained in many populations with relatively inexpensive and convenient technology. Some studies were devoted to determining appropriate models to represent the correlations between auditory stimuli and neural responses with EEG-based measures [e.g., 5-6]. One initial and key step of decoding continuous speech with EEG signals was to understand how the brain responds to the characteristics of auditory inputs [7].

With the EEG-based objective measures, the different functional roles of the auditory system can account for the effect of dynamic changes of speech features. Many studies used unnatural periodic stimuli (e.g., modulated tones, repeated phonemes, and discrete syllables) to investigate the neural functions of acoustic cues in the speech stimulus [e.g., 8]. Although those approaches with unnatural auditory stimuli provided valuable information to understand the neural mechanism of speech perception, there were differences in acoustical information and brain processed patterns existed between unnatural and continuous natural speech [9]. In order to explore the neural mechanisms underlying continuous speech processing, recent research showed that the cortical activities expressed the amplitude envelopes of natural speech at low frequencies indexed by the EEG signals [e.g., 6]. The regression methods, such as temporal response function (TRF), were widely used to describe the cortical tracking abilities to specific speech features (e.g., amplitude envelopes) [e.g., 4]. Some studies have shown that the correlations between neural activities and speech temporal envelops were affected by the auditory attentional modulation, speech intelligibility, and congruent visual interactions [e.g., 10-11]. These findings suggested that the cortical entrainment of speech amplitude envelops were jointly influenced by many functional roles with different neural populations for target speech processing. These analyses of cortical entrainment to speech stimuli were based on amplitude fluctuations of the entire auditory stimuli, while little evidence has been explored on how the low-frequency cortical entrainment to speech was affected with categorical speech segments.

The temporal and spectral discontinuities inside the speech signals were usually regarded as the speech landmarks, and the accurate identification of these cues provided large speech perception benefits in clean and noisy conditions [12]. Speech segments carrying different intelligibility information can be classified by the relative root-mean-square (RMS) intensity [13]. Some studies revealed that the different RMS-level segments had distinct effects on speech perception under clean and noise-masked conditions [e.g., 14-15], partly on account of speech features (e.g., vowels, consonants, and their transitions) with different intensities located at different RMS-level-specific speech segments. The various cortical tracking abilities were also found with the different RMS-level-based speech segments in clean conditions, indicating that the brain may process the natural ongoing speech as in categorical perception [16]. However, it is necessary to further understand

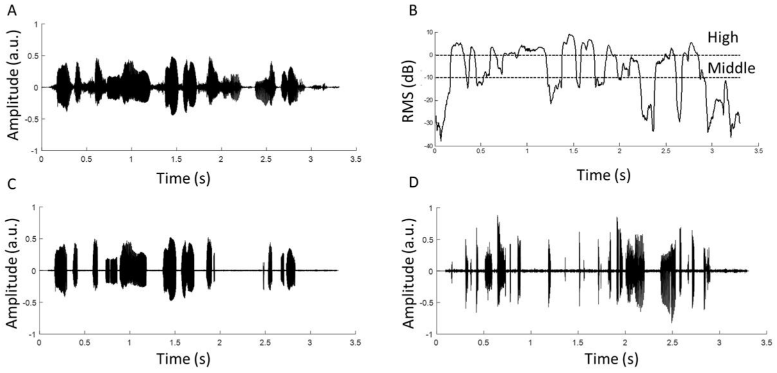

Figure 1. *A. The original speech of an example sentence. B. Relative root-mean-square (RMS) energy representations. The boundaries of different RMS-level segments were presented with the dashed lines. C. The sentence including only high-RMS-level segments. D. The sentence including only middle-RMS-level segments.*

the contribution of different RMS-level-specific speech segments to the target speech perception under conditions with various signal-to-masker ratios (SMRs).

The aim of this work was to investigate the contribution of high- and middle-RMS-level segments to target speech processing in different SMR conditions based on neural tracking activities. Since the TRF responses can analyze the cortical entrainments to speech envelope fluctuations at low frequencies, they were used in this study for evaluating the cortical responses to either high- or middle-RMS-level segments under 6, 3, 0, -3, and -6 dB SMRs. This work hypothesized that high- and middle-RMS-level segments yielded distinct relationships with low-frequency cortical oscillations reflected by the different morphological TRF responses. The target speech streams could evoke a larger TRF response than the ignored speech streams under conditions with high- or middle-RMS-level segments in a range of SMR conditions.

2. Methods

2.1. Participants

Twenty (12 males and 8 females, aged from 18 to 27 years) participants took part in this experiment. All participants had normal-hearing abilities (i.e., pure-tone threshold less than 25 dB at 125–8000 Hz). All subjects were native speakers of Mandarin Chinese and gave informed written consent prior to their participations. The experimental procedures were approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

2.2. Stimuli and experimental procedure

The stimuli used in this work were Maupassant's short fiction passages translated to Mandarin Chinese. Two of the four passages were read by a female speaker and the other two passages were read by a male speaker. To generate the auditory stimulus in each trial, the passages were divided into around 60-sec segments, with silent gaps less than 0.5 s. Mixed speech was formed with two speech streams from speakers of different genders. The RMS level of target speech was fixed, and the ignored speech was either the same or 6 dB, 3dB stronger/weaker than the target stream to form the 6, 3, 0, -3, and -6 dB SMRs.

The whole experiment was conducted in a double-walled acoustically shielded room. All auditory stimuli at a sampling rate of 16 kHz were presented bilaterally via headphones at 65 dB SPL using the software E-prime 2 [17]. All subjects were sat at a comfortable chair and looked at a fixation point on a computer screen. Each subject listened to a total of 100 trials at five SMR conditions, and each auditory stimulus only presented once. Prior to each block, a reminder was displayed on screen to direct the participant to focus on the female or the male speaker. Five trials with the same SMR were included in a block, followed by a three-minute break before the next block. During each break, the experimenter told participants the main idea of each block to avoid the effect of previous auditory stimuli on the next experimental trials.

2.3. Data analyses

Scalp EEG signals were recorded from 64 channels, with the sampling rate at 500 Hz. The extended 10-20 system was used to place scalp electrodes, with two additional reference electrodes at the bilateral mastoids. The electrode attached at the nose tip was served as the reference channel. Two electrodes placed above and below the left eye were used to measure the electrooculography (EOG) signals. All electrode impedances were kept under 5 kΩ. Participants were demanded to reduce body movements to avoid motion artifacts.

EEGLAB toolbox was implemented to process the offline EEG signals [18]. EEG data was digitally filtered from 1 to 50 Hz using a 4th-order Butterworth filter in both forwards and reversed directions to remove phase shifts. A 60-sec window was used to separate the filtered data into epochs, and each epoch included the whole duration of a trial. The independent component analysis was then conducted to remove the typical artificial components, such as the eye blinks, heartbeats and EOG. Subsequently, a 150th order zero-phase finite impulse filter was performed to digitally filter the continuous EEG data from 2 to 8 Hz, as cortical responses were phase-locked to speech envelopes in this low-frequency range.

The speech envelopes that represented the amplitude fluctuations at the low-frequency (i.e., 2–8Hz) were used in this study at the high- and middle-RMS-level speech segments. First, the sentences were divided into short-term segments using the Hamming windows with 16-ms block size and 50% overlap between adjacent windows. Then, the relative RMS-level was conducted to compute and classify the signal

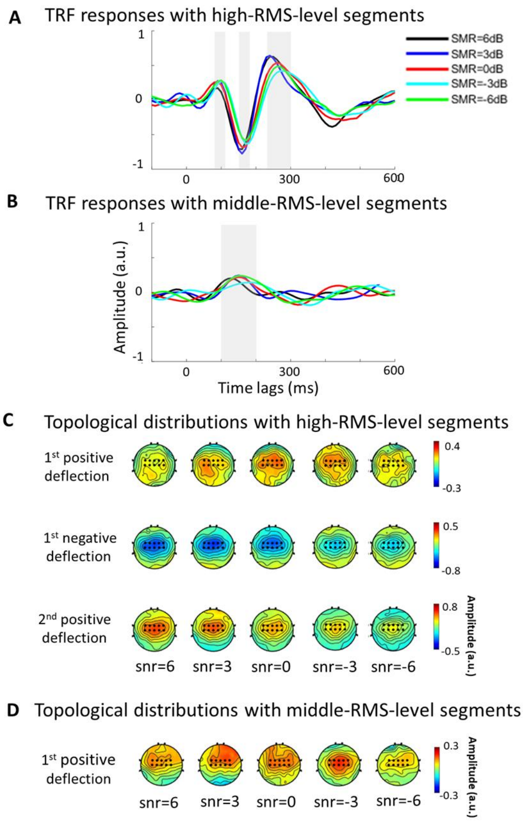

Figure 2. *A. Temporal response function (TRF) responses with high-RMS-level segments in different SMR conditions. B. TRF responses with middle-RMS-level segments in five SMRs. C. Topological distributions of TRF responses with high-RMS-level segments at three typical deflections. D. Topological TRF distributions with middle-RMS-level segments at a positive deflection.*

intensity in each windowed segment. As previous studies suggested [e.g., 14-15], the high-RMS-level segments were defined as the segments at and above the RMS level of the whole utterance (i.e., >0 dB in Fig. 1 C), while middle-RMS-level segments were the segments between 0 dB and -10 dB of the RMS level of the whole sentences (see Figure. 1 D). Figure 1 shows an original sentence and its high- and middle-RMS-level speech segments. Subsequently, the temporal envelopes of high- and middle-RMS-level segments were calculated by taking the absolute values of the Hilbert transform. Both the speech temporal envelopes and processed EEG signals were down sampled to 100 Hz, in order to reduce subsequent processing time.

The forward TRF model was used to measure the correlations between cortical tracking activities and natural speech stimuli at the high- or middle-RMS-level segments based on the description in the mTRF toolbox [18]. The regression procedure can be presented by the following function:

$$\mathrm{r(t,n)} = \sum_{\tau} w(\tau,n) * s(t-\tau) + \varepsilon(t,n) \qquad (1)$$

where $w(\tau,n)$ described the TRF responses for a specified range of time lags, τ, at the EEG electrode, n, to the auditory stimulus envelopes, $\mathrm{s(t)}$, at the instantaneous time, and $\varepsilon(t,n)$ is the residual response at channel n. The estimated neural response $\mathrm{r(t,n)}$ was calculated by a convolution of the stimulus envelopes $\mathrm{s(t)}$ with the specific TRF responses

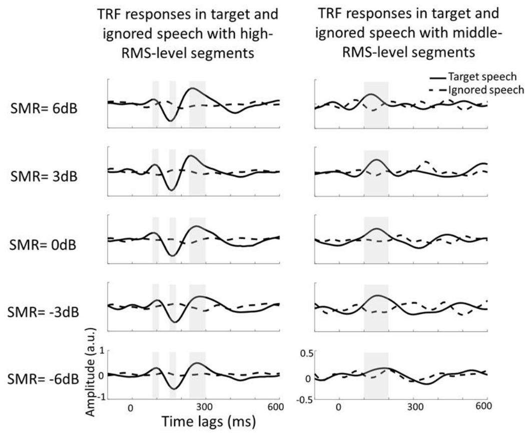

Figure 3. *The TRF waveforms to target and ignored speech streams calculated with high-RMS-level segments (left) and middle-RMS-level-segments (right) at five SMR conditions. The grey boxes show the time intervals of typical TRF responses.*

$w(\tau,n)$ at channel n. A leave-one-out cross validation approach was implemented to find the optimal ridge regression parameters by minimizing the mean-squared error between the original and estimated EEG responses. The grand-averaged TRF responses to high- and middle-RMS-level segments were computed across subjects in each SMR condition. The averaged TRF responses across active channels were calculated between -100 and 600 ms time lags. The topological distributions of TRF responses were displayed in typical time intervals corresponding to high- or middle-RMS-level segments. Furthermore, the TRF responses to the target and ignored speech streams were analyzed under the five SMR conditions with high- and middle-RMS-level segments, respectively.

3. Results

Figures 2 A and B display the grand averaged TRF responses across subjects at the electrodes located at the fronto-central regions. The topological distributions of high- and middle-RMS-level speech segments are shown in Figures 2 C and D with the active electrodes marked by black points under the five SMRs (i.e., 6, 3, 0, -3, and -6 dB). The grey windows present the time intervals of TRF responses with different TRF responses between target and ignored speech streams in high- and middle-RMS-level segments. Two positive deflections (i.e., 80–110 ms, and 230–300 ms) and a negative deflection (i.e., 150–180 ms) are used to show the TRF responses with high-RMS-level segments. The typical TRF responses of middle-RMS-level segments are displayed with a positive deflection from 100 to 200 ms time lags. Figure 3 shows the TRF response to target and ignored speech streams in high- and middle-RMS-level speech segments under five SMR conditions. TRF amplitudes and latencies under the significant time intervals were analyzed with high- and middle-RMS-level segments using the one-way repeated-measures analysis of variance.

3.1. Effect of different SMRs

As shown in Figure 2, with high-RMS-level segments, there are significant TRF amplitude differences at the first negative deflection (F (4, 76) = 5.162, p < 0.001) and the second positive deflection (F (4, 76) = 4.175, p = 0.004) in the five SMR conditions. Besides, the TRF latency shows a significant effect of SMRs at the first negative deflection (F (4, 76) = 7.567, $p < 0.001$) with high-RMS-level segments. The effect of SMR conditions, with middle-RMS-level segments, illustrates no significant differences in TRF amplitude and latency.

3.2. Effect of high- and middle-RMS-level segments

It is seen from Figure 3 that the target and ignored speech streams evoke different morphological TRF responses, with high- and middle-RMS-level segments. A negative and two positive deflections are displayed in high-RMS-level segments, while middle-RMS-level segments only show a positive deflection in TRF responses. When comparing the TRF responses between the target and ignored speech streams, the TRF responses in both high- and middle-RMS-level reflect significant differences under the typical time intervals in most SMR conditions with the sample-to-sample t-test ($p < 0.05$; excluding the test condition of -6 dB SMR with middle-RMS-level segments).

4. Discussion and conclusions

The aim of the present study was to explore the effects of high- and middle-RMS-level segments on cortical tracking activities to target speech in different SMR conditions using TRF responses. The experimental results were consistent with previous studies [e.g., 8], indicating that the low-frequency cortical activities could track the target speech envelopes in a range of SMR. In addition, different speech segments employed distinct cortical response patterns to achieve the successful recognition of target speech with the competing interference [e.g., 6, 19]. In this study, the distinct morphological distributions of the TRF responses to high- and middle-RMS-level segments indicated that the brain responses to natural speech could be varied in speech segments based on the different RMS intensities. Although the neural activities showed a weaker correlation to middle-RMS-level segments than those to high-RMS-level segments, the middle-RMS-level segments could be used to identify the target speech stream by comparing the cortical responses of the target stream with those of the ignored stream. Hence, the TRF results in this study suggested that high- and middle-RMS-level segments played distinct and important roles in decoding target speech with the corresponding EEG signals. The effect of different RMS-level segments should be considered in further studies to improve the performance of continuous speech decoding.

In conclusion, the current EEG-based study illustrated that both high- and middle-RMS-level speech segments contributed to the target speech perception in a range of SMR conditions. Even with the decrease of SMR, the TRF responses showed more robust and significant patterns to high- and middle-RMS-level segments of the target speech than those of the ignored speech. Besides, TRF morphologies were also different under the test conditions with high- and middle-RMS-level speech segments. These results suggested that the relative RMS intensities inside continuous speech could be a crucial factor for understanding target speech decoding mechanisms in noisy conditions.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61971212), the Basic Research Foundation of Shenzhen (Grant No. GJHZ20180928155002157), and High-level University Fund G02236002.

6. References

[1] Cherry, E.C., "Some experiments on the recognition of speech, with one and with two ears," *J Acoust Soc Am.,* vol. 25, pp. 975–979, 1953.

[2] Ding, N., and Simon, J. Z., "Emergence of neural encoding of auditory objects while listening to competing speaker,". *Proc. Natl. Acad. Sci. U. S. A.,* vol. 109, pp. 11854–11859, 2012.

[3] Mesgarani, N., and Chang, E.F., "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature,* vol. 485, 233–236, 2012.

[4] O'sullivan, J.A., Power, A. J., Mesgarani, N., Rajaram, S., and Lalor, E. C., "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cereb. Cortex,* vol. 25, no. 7, pp. 1697–1706, 2014.

[5] Horton, C., Srinivasan R., and Zmura, M. D, "Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'," *J. Neural Eng.,* vol. 11, no. 4, pp. 046015, 2014.

[6] Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C., "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Curr. Biol.,* vol. 25, pp. 2457–2465, 2015.

[7] Wu, M., David, S. V., and Gallant, J. L., "Complete functional characterization of sensory neurons by system identification," *Annu. Rev. Neurosci.,* vol. 29, pp. 477–505, 2006.

[8] Anderson, S., Parbery-Clark, A., White-Schwoch, T., and Kraus, N., "Auditory brainstem response to complex sounds predicts self-reported speech-in-noise performance," *J. Speech Lang. Hear. Res.,* vol. 56, no. 1, pp. 31–43, 2013.

[9] Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F., "Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli," *J. Neural Eng.,* vol. 36, no. 6, pp. 2014–2026, 2016.

[10] Biesmans, W., Das, N., Francart, T., and Bertrand, A., "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE T. Neur. Sys. Reh.,* vol. 25, no. 5, pp. 402–412, 2017.

[11] Teoh, E. S., and Lalor, E. C., "EEG decoding of the target speaker in a cocktail party scenario: considerations regarding dynamic switching of talker location," *J. Neural Eng.,* vol.16, no. 3, pp. 036017, 2019.

[12] Li, N., and Loizou, P. C., "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *J. Acoust. Soc. Am.*, vol. 124, no. 6, pp. 3947–3958, 2008.

[13] Kates, J. M., and Arehart, K. H., "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.,* vol. 117, no. 4, pp. 2224–2237, 2005.

[14] Chen, F., and Loizou, P. C., "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4104–4113, 2012.

[15] Chen, F., and Wong, L. L., "Contributions of the high-RMS-level segments to the intelligibility of mandarin sentences," Proc. ICASSP, pp. 7810–7814, 2013.

[16] Wang, L., Li, H., Wu, E. X., and Chen, F., "Cortical auditory responses index the contributions of different RMS-level-dependent segments to speech intelligibility," *Hearing Res.,* vol. 383, pp. 107808, 2019.

[17] Schneider, W., Eschman, A., and Zuccolotto, A., "E-Prime: User's guide," *Psychology Software Incorporated*, 2002.

[18] Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C., "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers Hum. Neurosci.,* vol. 10, no. 604, 2016.

[19] Oganian, Y., and Chang, E. F., "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Sci. Adv.*, vol. 5, no.11, pp. 6279, 2019.