Cortical Entrainment to Speech Produced by Cochlear Implant Users and Normal-Hearing Talkers

Shruthi Raghavendra, Sungmin Lee, Fei Chen, Senior Member, IEEE, Brett A. Martin, and Chin-Tuan Tan, Member, IEEE

Abstract—The perceived sound quality of speech produced by hard-of-hearing individuals greatly depends on the degree and configuration of their hearing loss. A cochlear implant (CI) may provide some compensation and auditory feedback to monitor/control speech production. However, to date, the speech produced by CI users is still different in quality from that produced by normal-hearing (NH) talkers. In this study, we attempted to address this difference by examining the cortical activity of NH listeners when listening to continuous speech produced by 8 CI talkers and 8 NH talkers. We utilized a discriminative model to decode and reconstruct the speech envelope from the single-trial electroencephalogram (EEG) recorded from scalp electrode in NH listeners when listening to continuous speech. The correlation coefficient between the reconstructed envelope and original speech envelope was computed as a metric to quantify the difference in response to the speech produced by CI and NH talkers. The same listeners were asked to rate the perceived sound quality of the speech as a behavioral sound quality assessment. Both behavioral perceived sound quality ratings and the cortical entrainment to speech envelope were higher for the speech set produced by NH talkers than for the speech set produced by CI talkers.

I. INTRODUCTION

Quality measures assess “how” a speaker produces an utterance [1]. Judgement of sound quality is unique to each listener because each listener has a unique internal reference for determining the degree of degradation and distortion (in this context, speech produced by deaf/hard-of-hearing individuals) when making a perceived quality judgment of the speech. The degree and configuration of hearing loss in a talker may affect the way the speech is produced, and consequently, affect the quality as perceived by a listener. Cochlear implants (CI) may help individuals with hearing loss to restore or improve the ability to hear and, provide the auditory feedback necessary to produce speech. However, there remains a large variability in speech production proficiency and intelligibility among implant recipients which may be complicated by factors such as the age of implantation, duration of hearing loss, the presence of residual hearing, etc. [2–4]. The aim of this study was to investigate the processes underlying these quality judgments using electroencephalography (EEG) while listeners heard speech differing in quality.

Short stimuli (i.e., clicks, tone pips, and phonemes) have been classically used to study the effect of various degradations (for instance, noise and distortion) on speech discrimination with event-related potentials (ERPs) to elicit P3, a positive peak that appears in the vicinity of 300 msec after the onset of a target stimulus. In general, as the level of degradation of the stimuli decreased, the harder it was for the subjects to discriminate them, resulting in lower amplitude and longer latency of the P3 component [5–9]. This technique may not apply when sound quality judgment is made on continuous speech, where consecutive speech tokens (e.g., phonemes) are overlapped and dynamically changing over time [10].

Cortical entrainment to the speech envelope may provide a useful tool to investigate the neurophysiologic processing of continuous speech as the dynamic cortical activity tracks the envelope of continuous, natural speech [11]. It has been demonstrated that the speech envelope can be decoded from single-trial EEG/MEG recordings obtained by presenting the stimulus only once [12]. The temporal envelope, a slow variation of amplitude of speech (<10 Hz) [12] is considered to be one of the most important cues for speech intelligibility [13] and speech perception [14]. Recent work by [15] proposed a multivariate linear discriminative model that maps the multi-channel EEG signal into a single-channel speech envelope. Using this technique, it has been shown that the speech envelope can be decoded from single-trial EEG/MEG recordings obtained by presenting the stimulus only once [16–18]. Previous research used this technique to investigate the cortical entrainment to continuous speech envelope fluctuations in response to speech intelligibility and showed that higher speech intelligibility coincided with improved cortical entrainment to the speech envelope in normal-hearing (NH) listeners [16–18]. Likewise, we used this method to study the difference in the cortical entrainment to the speech envelope in relation to the perceived sound quality, as rated by NH listeners, between the speech stimulus set spoken by CI talkers and NH talkers.

II. MATERIALS AND METHODS

Figure 1 shows the overall setup for both behavioral and electrophysiological experiments with NH participants listening to the speech produced by both CI talkers and NH talkers. The behavioral experiment was conducted first followed by the EEG experiment. In the behavioral experiment, each listener was asked to rate the perceived sound quality of a stimulus set consisting of 16 speech passages spoken by 8 CI talkers and 8 NH talkers. In the...
electrophysiological experiment, the cortical responses of the same listeners were recorded using EEG while listening to the same speech passages presented in the behavioral experiment. For each speech passage, the speech envelope was reconstructed from its associated EEG signal using a decoder and the bootstrapped Spearman correlation between the actual speech envelope and the reconstructed envelope was employed as a metric that measured the cortical entrainment to the actual speech envelope in each NH listener.

A. Participants

11 NH adult listeners, who were native speakers of American English and aged from 19 to 29 years (mean age = 21.5 years; 5 female, 6 male), participated in this study. All NH listeners were screened by presenting pure tones at 20 dB HL from 250 Hz to 8 kHz at octave frequencies and had normal hearing thresholds <20 dB HL. Experiments were conducted following the protocols approved by the Institutional Review Board at the University of Texas at Dallas. The informed consent forms were obtained from the participants before the experiment. Their participation was paid in the study.

B. Stimuli

Speech passages used in this study were extracted from the “Corpus of deaf speech for acoustic and speech production research” database collected at the University of Memphis [19]. This corpus is a pool of speech recordings digitally sampled at 44100 Hz, with hearing-impaired (HI) talkers and NH talkers reading the entire ‘The Rainbow Passage’ (using Shure SM93 prologue dynamic microphone). We selected 8 passages read by CI talkers from the HI group (4 female; 4 male) and 8 passages from the NH group (6 female; 2 male) to form two stimulus sets that were referred to as stimuli CI set (stimulus_CI1 to stimulus_CI8) and stimuli NH set (stimulus_NH1 to stimulus_NH8) respectively throughout the study.

C. Behavioral experiment

Before the experiment started, listeners were asked to sit comfortably in a soundproof booth in front of a touch screen computer monitor. They were positioned 1 m from a loudspeaker at 0° azimuth at ear height. The speech passages were normalized for RMS (root mean square) amplitude and presented via the frontal loudspeaker at an intensity of 65 dB SPL. A listener was asked to perform two trials of behavioral sound quality assessment i.e., in one trial, the 16 speech passages (stimuli_CI and stimuli_NH sets) were presented one at a time in a randomized order to a listener. Each stimulus was presented only once. Listeners were instructed to listen to the speech passages and to rate the sound quality of each passage on a Likert 10 point scale [20], with 1 being the most distorted and 10 being the most undistorted using a touch screen monitor. In the second trial of behavioral sound quality assessment, the same listener was asked to rate the perceived sound quality of the passages again presented in a randomized order from the first trial. Finally, the perceived sound quality rating of each passage was obtained as the average of the ratings across the two trials.

D. EEG experiment

EEG recordings were obtained after the behavioral sound quality rating assessment. A 64-channel actiCHamp amplifier EEG setup (Brain Products GmbH, Munich, Germany) was used to record the ongoing single-trial EEG in response to the same set of stimuli (stimuli_CI and stimuli_NH sets) used in the behavioral test (presented in random order). The EEG signals were recorded using an electrode cap (actiCAP, Brain Products GmbH, Munich, Germany) placed on the listener’s scalp in accord to 10-20 system [21]. Each listener was asked to minimize body movement and watch a silent movie with captions on a monitor, while EEG recording (sampling rate = 1 kHz) was in progress.

E. Signal processing: EEG and speech

EEG data were preprocessed using the EEGLAB toolbox [22] for MATLAB to prune unwanted artifacts. The portion of the EEG contaminated with artifacts related to muscle was removed by visual inspection. Artifacts from the eye blink, lateral eye movement, and heartbeat were pruned from the EEG using independent component analysis (ICA) in the EEGLAB toolbox. The preprocessed EEG signal was filtered from 0.1 to 10 Hz [12] using a 6th-order Butterworth filter. Then EEG data were downsampled to 128 Hz to reduce computation time. The speech envelope was extracted from the speech using the Hilbert transformation approach. The speech envelope was filtered between 0.1 and 10 Hz using a 6th-order Butterworth filter was downsampled to 128 Hz.

F. Speech envelope reconstruction

A linear decoder as proposed in [15] was used to predict and reconstruct the speech envelope from the associated EEG activity. A decoder acts as a spatiotemporal filter that linearly maps the EEG to the speech envelope thereby reconstructing the speech envelope estimated from the corresponding EEG response recorded when listening to the speech stimulus. For this, the time-shifted version of the single-trial EEG channels was obtained by applying a range of delays (e.g., 0 and 500 ms) to each channel, and then all of the delayed channels were weighted, in order to linearly reconstruct the envelope. The actual speech envelope and the reconstructed envelope were then correlated with each other, which yielded a measure of cortical entrainment to the actual speech envelope. The process is explained as follows: Given a linear decoder \( g \) that integrates the lagged time series of EEG response \( r \) over an integration window length \( \tau \), a single estimate of the stimulus envelope \( \hat{s}(t) \) is computed as:

\[
\hat{s}(t) = \frac{1}{T} \sum_{\tau} R(t + \tau, n) g(\tau, n)
\]

with \( t \) ranges from 0 to T. \( R \) represents the lagged time series of EEG response \( r \) and \( n \) is the index of the EEG channel.
The decoder $g$ is derived by minimizing the mean-squared-error (MSE) between the original speech envelope and the reconstructed envelope, i.e.,

$$g = \arg \min (|\delta(t) - s(t)|^2)$$  \hspace{1cm} (2)

where $E$ represents the expected value, and $s(t)$ and $\delta(t)$ denote the original speech envelope and the reconstructed envelope respectively. The decoder $g$ is computed using the ridge regression by solving the equation below:

$$g = (R'R)^{-1}R's$$  \hspace{1cm} (3)

where the superscript $T$ represents the transpose of a matrix. are overlapped and dynamically changing over time [10].

Figure 2 illustrates the procedure of reconstruction of the speech envelope in a single listener. For each NH listener, 16 decoders were built using each of the 16-speech envelope-EEG pairs respectively as in Eq. (3). To reconstruct the envelope which corresponds to stimulus_CI1, ‘Average decoder’ was first computed by averaging the remaining 15 decoders, i.e., decoder 2 to decoder 16 and convolved with the EEG signal recorded during the presentation of stimulus_CI1 (Eq. (1)). This ‘leave-one-out’ model approach was repeated to reconstruct the envelope for each of the remaining 15 speech passages for the same listener.

Finally, the bootstrapped Spearman correlation was computed between the actual speech envelope and the reconstructed envelope as the metric to measure the degree of cortical entrainment to the actual speech envelope. For this, we randomly permuted the reconstructed envelope 1000 times and calculated Spearman’s correlation between the result and the actual speech signal for each permutation. The final correlation value was evaluated as the averaged value of the resulted 1000 correlation values. A higher correlation value between the actual speech envelope and the reconstructed envelope indicates higher cortical entrainment of the actual speech envelope.

III. RESULTS

A. Behavioral results

Figure 3 presents the mean and variance of the perceived sound quality ratings for each stimulus in stimuli_CI and stimuli_NH sets across all NH listeners. The perceived sound quality ratings within the stimulus_CI set varied widely on the scale; stimulus_CI2 had the highest mean rating of 9.2 and stimulus_CI8 had the lowest mean rating of 1.2 across all the listeners. For the stimulus_NH set, there was little difference in ratings among stimuli as each was rated almost equally high in perceived sound quality.

Statistical analysis shows that the group mean of the perceived sound quality rating for the stimulus_NH set (mean: 9.4, SD: 0.8) was higher than that of stimulus_CI set (mean: 6.9, SD: 3). A paired-samples t-test was performed on the perceived sound quality ratings between the stimulus_CI and stimulus_NH sets to investigate whether the difference in perceived sound quality ratings between two stimulus sets was significant or not. The results showed that the perceived sound quality ratings for the stimulus_CI set were significantly lower than those of the stimulus_NH set ($t(87) = -7.4$, $p < 0.001$).

B. EEG results

Figure 4 presents the cortical entrainment to speech envelope in both stimulus sets across NH listeners. In the stimulus_CI set, stimulus_CI2 was rated with the highest mean perceived sound quality rating of 9.2 and held the highest median cortical entrainment (0.03) across the listeners. A similar observation was not found in the stimulus_NH set. Apparently, within the stimulus_NH set, there was no evident difference in the perceived sound quality ratings among the listeners. The group mean Spearman correlation value for the stimulus_CI set was -0.004 (SD: 0.06) and was 0.01 for stimulus_NH set (SD: 0.06) across the listeners. A paired-samples t-test performed on the cortical entrainment to speech envelope between the two stimulus sets in question also revealed a significant ($p < 0.04$) difference in cortical entrainment between the two stimulus sets ($t(87) = -2.07$). The averaged cortical entrainment for the stimulus_NH set was statistically higher than the equivalent for the stimulus_CI set. In general, there was more variability in the median entrainment values within the stimulus_CI set as compared to the stimulus_NH set.
IV. DISCUSSION AND FUTURE WORK

The method applied in this study was previously used to successfully predict speech intelligibility [18] and auditory attention detection [23] from the cortical entrainment of the speech envelope. To our knowledge, the literature lacks a study of cortical entrainment of the speech in relation to perceived sound quality based on a single-trial of EEG data. This study lays the foundation by presenting an algorithm which shows that the perceived sound quality difference observed between the two stimulus sets by NH listeners is also reflected in the cortical entrainment of the speech envelope in the same listeners. We did not attempt to correlate the perceived sound quality ratings with the cortical entrainment of speech envelope and this is because the behavioral data were collected when listeners were actively listening to speech passages, whereas EEG data were obtained when participants were listening passively.

This preliminary work employed a linear model to predict and reconstruct the speech envelope from the EEG signal. A simple linear decoder is probably not a good fit as the neurons are known to respond to a complex stimulus like speech in a non-linear manner [24]. Therefore, our future work would include a non-linear decoder in an attempt to reconstruct the speech feature more accurately from the cortical responses. Also, the future study would consider the inclusion of other speech features such as spectrogram and phoneme-related features to investigate the relation of cortical entrainment to these features to the perceived sound quality.

REFERENCES

[1] P. C. Loizou, “Speech quality assessment,” In Multimedia analysis, processing and communications, pp. 623-654, Springer, 2011.
[2] H. Kim et al, “Cochlear implantation in postlingually deaf adults is time-sensitive towards positive outcome: prediction using advanced machine learning techniques,” Sci Rep. 2018;8(1):18004, 2018.
[3] S. Ruff, T. Bocklet, E. Nöth, J. Müller, E. Hoster, M. Schuster, “Speech production quality of cochlear implant users with respect to duration and onset of hearing loss,” ORL, 79(5), 282-294, 2017.
[4] A. Gautam, J. G. Naples, S. J. Eliades, “Control of speech and voice in cochlear implant patients,” The Laryngoscope, 129(9), 2019.
[5] B. A. Martin, A. Sigal, D. Kurtzberg, D. R. Stapells, “The effects of decreased audibility produced by high-pass noise masking on cortical event-related potentials to speech sounds /ba/ and /da/,” J Acoust Soc Am, 101, pp. 1585- 1599, 1997.
[6] B. A. Martin and D. R. Stapells, “Effects of low-pass noise masking on auditory event-related potentials to speech,” Ear Hear, 26, 195–213, 2005.
[7] J. N. Antons, A. Porbadnigk, R. Schleicher, B. Blankertz, S. Möller, G. Curio, “Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise,” In: Proceedings of the audio engineering society (AES), 2010.
[8] J. N. Antons, R. Schleicher, S. Arndt, S. Möller, A. K. Porbadnigk, G. Curio, “Analyzing speech quality perception using electroencephalography,” IEEE J. Sel. Top. Signal Process. 6, 721–731, 2012.
[9] A. K. Porbadnigk, M. S. Treder, B. Blankertz, J. N. Antons, R. Schleicher, S. Möller et al., “Single-trial analysis of the neural correlates of speech quality perception,” J. Neural Eng. 10:056003, 2013.
[10] D. Lesenfants, J. Vanthornhout, E. Verschueren, L. Decruy, T Francart, “Predicting individual speech intelligibility from the neural tracking of acoustic-and phonetic-level speech representations,” bioRxiv, p.471367, 2018.
[11] S. J. Aiken and T.W. Picton, “Human cortical responses to the speech envelope,” Ear Hear,29, 139–157, 2008.
[12] N. Ding and J. Z. Simon, “Cortical entrainment to continuous speech: functional roles and interpretations,” Front Hum Neurosci 8:311, 2014.
[13] J. E. Peelle and M.H. Davis, “Neural oscillations carry speech rhythmthrough to comprehension,” Front Psychol 3:320, 2012.
[14] R. V. Shannon, F-G. Zeng, V. Kamath, J. Wygonsk, M. Ekelid M, “Speech recognition with primarily temporal cues,” Science, 270 (5234) (1995), pp. 303-304, 1995.
[15] M. J. Crosse, G. M. Di Liberto, A. Bednar, E. C. Lalor, “The multivariate temporal response function (mTRF) Toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli,” Front Hum Neurosci, 10, 604, 2016.
[16] N. Ding, J. Z. Simon, “Adaptive temporal encoding leads to a background-insensitive cortical representation of speech,” J Neurosci 33(13):5728–5735, 2013.
[17] Y. Y Kong, A. Somarowthu, N. Ding, “Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech,” J Assoc Res Otolaryngol 16(6):783–796, 2015.
[18] J. Vanthornhout, et al., “Speech intelligibility predicted from neural entrainment of the speech envelope,” JARO - Journal of the Association for Research in Otolaryngology, pp 1–11, 2018.
[19] L. L. Mendel, S. Lee, M. Pousson, C. Patro, S. McSorley, B. Banerjee, S. Najnin, M. H. Kapoorchal, “Corpus of deaf speech for acoustic and speech production research,” The Journal of the Acoustical Society of America 142, EL102, 2017.
[20] M. Sangthong, “The effect of the likert point scale and sample size on the efficiency of parametric and nonparametric tests,” Thailand Statistician;18(1): 55–64, 2020.

[21] R. Oostenveld, P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,” Clin. Neurophysiol., 112, pp. 713-719, 2001.

[22] A. Delorme, S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” J. Neurosci. Methods, 134, pp 9–2, 2004.

[23] J. A. O’Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinnecunningham, M. Slaney, S. A. Shamma, E. C. Lalor, “Attentional selection in a cocktail party environment can be decoded from single-trial eeg,” Cereb Cortex 25(7):1697–1706, 2014.

[24] F.E. Theunissen, S.V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L, “Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli,” Network 12, 289–316, 2001.