Neural Tracking of Band-Limited Sine-Wave Speech in Normal Hearing and Cochlear Implant Listeners

Sungmin Lee[1,*], Sara Akbarzadeh[2] and Chin-Tuan Tan[2]

[1]Tongmyong University, 428 Sinseon-ro, Nam-gu, Busasn, 45820, Republic of Korea

[2]Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, 800 West Campbell Road, Richardson, TX, 75080, USA

Abstract: A sine-wave speech represents complex speech with a limited number of sinusoidal components. Functionally, its strategy draws a close similarity to the signal processing strategy in cochlear implant (CI) which transmits speech envelope information in limited number of channels. In this study, we investigated how synchronous cortical activities to speech envelope relates to the speech intelligibility of sine-wave speech in different bandwidths with both normal hearing (NH) and CI listeners. 12 NH and four CI participants were recruited. We divided our NH participants into two groups: 1) Six NH listeners, 2) six NH listening to CI simulation synthesized using a noise vocoder. In our third group, 3) four CI users, one of them participated in only behavioral tests. Neural tracking was obtained using multi-channel electroencephalogram (EEG) system and cross-checked with behavioural performance, speech perception scores and speech quality ratings. Our result showed that intelligible sine-wave speech can be built with a small number of sinusoidal components selected from the original speech spectrum. Our cross-correlation analysis between cortical activities and speech envelope fluctuation showed an increasing trend in their synchrony with sine-wave speech built using more sinusoidal components for both group 2) and 3). Cortical entrainment to speech envelope is more observable in CI users compared to NH listened to CI simulated sine-wave speech. Our result have implications for understanding of neural tracking in terms of spectrally degraded speech perception in NH and CI individuals.

Keywords: Sinusoidal model, electroencephalogram (EEG), cortical entrainment, speech intelligibility, hearing prosthesis.

INTRODUCTION

Sinusoidal Model (SM)

Speech produced from human vocal systems has acoustically complex nature caused by the mixture of variances in vibrations of vocal folds, resonance frequency of vocal tract, and dynamic of articulators. However, some findings from signal processing studies have reported that the vast amount of complex speech signal is full of redundancy, and can be perceptually represented with few speech information [1-4]. One of the approaches is the sinusoidal synthesis and analysis model (SM) proposed by McAulay and Quatieri [5] which represents the complex speech signals with a limited number of sinusoidal components. In the SM analysis, amplitude, phase and frequency parameters are extracted from input speech signals using a short-time fast Fourier transform (STFT) at each time frame. Then, prominent amplitude peaks of resulting frequency spectra are identified. Spectral tracks are created by linking the spectral peaks of adjacent frames which occur at similar frequencies. Cubic phase interpolations are then used to maximally smooth the phase track and eliminate the phase discontinuity. The output speech is eventually resynthesized using only sinusoidal components on the spectral tracks.

In general, eliminating redundant elements and retaining only a few sinusoidal components in speech may reduce the intelligibility of speech with normal hearing (NH) listeners. Kates [6] investigated whether the SM can be used as a noise reduction technique to improve signal to noise ratio (SNR) when the level of speech is higher than that of background noise. Speech with only 8 and 16 sinusoidal components were presented to NH listeners in consonant recognition and perceived speech intelligibility tasks. The results showed lower performance for the SM speech (8 components: 73%, 16 components: 83%) in comparison with the original speech (92%). They showed that the SM may not serve effectively as a speech enhancement technique in noise conditions for NH listeners. However, the intelligibility of speech was still largely retained with a limited number of sinusoidal components. Timms [7] investigated the effect of spectral components in the SM by varying the window length using German vowels, consonants and sentences perception. They reported that the different number of spectral components are needed for NH listeners based on the given different temporal/spectral resolution to reach 100% sentence recognition. For example, at least two to four sine wave components per 1.5 ms were necessary to achieve 100%. The

*Address correspondence to this author at the Tongmyong University, 428 Sinseon-ro, Nam-gu, Busasn, 45820, Republic of Korea; Tel: 82-51-629-2134; Fax: 82-51-629-2019; E-mail: slee18@tu.ac.kr

E-ISSN: 2311-1917/20

notable consistency between two studies is that only a few spectral sinewave components enable to achieve considerably high speech perception.

Application of SM to Cochlear Implant (CI)

One possible application of spectrally reduced speech is CI, which is a prosthetic device to help people with severe to profound sensorineural haring loss. Unlike NH which uses peripheral (outer, middle, and inner ears) and central auditory mechanisms, CIs bypass many of these structures and directly stimulate the auditory nerve fibers with electrical pulses via an array of electrodes implanted in the cochlea. In the speech processing of CI, incoming acoustic speech signals are filtered into channels which are associated with intra-cochlear electrodes to mimic the tonotopicity in cochlea by having high frequency channel assigned the basal electrode and low frequency channel assign to apical electrode. The envelope of each channel output is extracted and used to modulate biphasic electrical pulses which is delivered as the electrical stimulation to auditory nerves via the associated intra-cochlear electrode. Regardless of different speech processing strategies and frequency-to-channel mappings deployed to deliver the signal, one common fact of electrical hearing by CI is that it uses a limited number of frequency channels (typically 8 - 22 electrodes). Spectrally reduced sound is perceptually unnatural to acoustic hearing ears of healthy hearing individuals. However, CI is able to deliver intelligible speech with limited number of channels to the patients with a great benefit. Speech processing in CI seems to draw a similar scheme as the SM in the way that both strategies select limited, but important, speech information among complex speech signals. This study investigated the perceptual performance and neural correlation of speech processed by SM which is functionally analogous to CI system.

Cortical Entrainment to Speech Envelope

In recent years, the phenomenon of cortical entrainment to speech envelope has received much attention in a neuroscience field where neural representations are explored in connection with speech perception mechanism. In cortical entrainment studies, scientists were presenting continuous speech instead of short transient segment of speech used in the traditional event-related potentials, while recording the corresponding neural synchrony. Compared to the traditional approaches, this approach of tracking synchronous cortical activity to the realistic dynamics of continuous speech may better reflect the cortical process in perceiving speech. Previous studies [8-11] had demonstrated robust cortical entrainments to continuous speech stream. The cortical synchrony has been found to be significantly enhanced with speech intelligibility of listeners [12-14], and attention of listeners [14-16]. Cortical activity has been found to be synchronized not only with hierarchical linguistic structures, such as words, phrases and sentences [18,19], but also with smaller units of speech features, such as syllable [12]. Some studies, however, showed that cortical entrainment appears to be more associated with acoustic changes, rather than implying linguistic processing for speech recognition [20].

To our knowledge, only few studies had worked on this topic with CI's. Kong *et al*., [21] used a number of noise vocoders, which are analogous to CI simulations with NH listeners, to examine the effect of spectral degradation on selective attention. They reported that attentional modulations derived from cortical entrainment to noise vocoded speech are more robust for the larger number of channels than those for the less number of channels, implying detrimental effect of degraded sensory input (similar to CI stimulation) for neural stream segregation when a mixture of speech are presented. Verschueren *et al*. [22] first revealed the neural tracking of the speech envelope in CI users with a successful artifact-rejection method, which periodically leaves out small groups of stimulation pulses that may be timely associated with electric artifacts occurred from CI. Using the same artifact rejection technique, they further investigated the neural tracking phenomenon in CI users. They varied electrical presentation levels by changing current unit (cu) and corresponding electroencephalogram (EEG) responses were examined in relation to CI users' speech perception scores. The study found that cortical synchrony to speech and speech perception ability improves with increasing stimulation level. Higher correlation between neural tracking and speech intelligibility that is typically found in NH listeners, was observed in CI users. Outcomes of these studies support that the neural envelope tracking can potentially be an objective measure of CI performance.

In our study, we used a cross-correlation function to analyse the synchrony between cortical activities and speech envelope. The cross-correlation procedure is a common approach used to track similarities between two series of streams in time using a varying lag windows. The best time lag/frame where EEG coefficients and speech envelope matched up was identified in the analysis. According to Aiken and Picton [13] the cortical activities begin to follow speech envelope at delays ranging from 150 to 200 msec. Kumagai *et al*. [23] used the cross-correlation to investigate the effect of familiarity of music based on the degree of cortical entrainment. Two prominent peaks of the cross-correlation were identified at the time lags around 70 and 140 msec, and those peaks were larger for unfamiliar music than familiar music implying stronger neural activity in response to unfamiliar sound. Given the fact that CI mainly utilizes envelope information in speech, a metric to quantify cortical entrainment to speech envelope will make a useful measure to understand the underlying neural activities in perceiving speech by the CI population.

Aim of Study

This study examined the speech intelligibility of sine-wave speech with different number of sinusoidal components by limiting the frequency range of sine-wave speech with selected bandwidth (BW). We hypothesized that neural tracking to speech envelope will provide a more central metric to measure the intelligibility of sine-wave speech, and the associated number of sinusoidal components needed to facilitate the intelligibility. Higher degree of synchrony (represented by higher cross-correlation value) between cortical activities and sine-wave speech envelope is anticipated when spectra-temporal structure of sine-wave speech is able to facilitate higher speech intelligibility. Sine-wave speech built with the larger number of spectral components and wider BWs are expected to be more intelligible under the hypothesis.

METHODS

Participants

Twelve NH subjects aged from 18 to 28 [M = 23.33, SD = 3.19], and four CI users aged from 23 to 65 [M = 43, SD = 23] participated in the current study. They all are native speakers of American English with no history of mental illness and cognitive problem. For NH group, their normal audibility (thresholds < 25 dB HL) was verified by presenting pure-tones at 20 dB HL using a head phone at each ear. Aided and un-aided pure tone audiometry were conducted to identify hearing thresholds of CI users. All CI participants had severe-to-profound sensorineural hearing loss, showing the aided threshold levels better than 35 dB HL when compensated by their CI devices (Nucleus 6 at Cochlear America). Three subjects were pre-lingual deafness, but one had hearing loss after language acquisition. Two of them were bilateral and the other two used CI in either left ear or right ear. Etiology varies for each subject (noise induced, ototoxicity, genetic, and meningitis) and their primary use of communication mode is oral and lip reading. We also employed a method (AngelSim V1.08.01) to simulate an 'equivalent' CI stimulation for NH listeners by pre-processing speech using eight-channel noise vocoder with a frequency range between 200 Hz – 7 kHz (24 dB/octave of filter slope). Six of the 12 NH subjects were asked to listen to the CI simulated speech.

For the whole experiment, our subjects were divided into three listening groups: 6 NH listeners listen to SM speech (control group), 6 NH listeners listen to SM speech processed with the noise vocoder (CI simulation), and 4 CI listeners listen to sine-wave speech. Subjects were paid for their participation. The experimental protocol employed was approved by The University of Texas at Dallas Institutional Review Board.

Stimuli

The material used to measure intelligibility of sine-wave speech was AzBio sentence list [24]. It was developed for the purpose of speech perception evaluation for CI patients. Twelve out of the 15 AzBio lists were randomly chosen to be processed or resynthesized from 12 conditions by having different combinations of the three numbers of sinusoidal components (1, 2, and 6) and four BWs (cut-off frequency of low-pass filters at 1 k, 1.5 k, 3 k, and 6 kHz). "Sinewave and Sinusoid+Noise Analysis/ Synthesis" model (Ellis, 2003) in MATLAB was used to resynthesize the sine-wave speech from limited number of sinusoidal components. In the model, STFT was implemented with 256-point Fast Fourier Transform (FFT) and windowed data segments of 256 samples (5.8 ms) sliding at 128 samples (2.9 ms) for an overlap of 50%.

$$STFT\ [x(n)](m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x_n w_{n-m} e^{-i\omega n} \qquad (1)$$

In STFT (see Equation 1), x_n and w_n indicate the input signal to be transformed and the window function, respectively, in time domain. The Fourier transform of the signals that represents the phase and magnitude over time (m) and frequency (ω) is denoted by X. At each time frame, the spectral components that peak from the spectrum in magnitude were selected. The selected peaks in adjacent frames, whose frequencies are in close proximity, were linked to create spectral tracks and smoothed using a quadratic interpolation in the second pass of the algorithm (see Figure **1A**). The spectral components remained on the spectral tracks are considered as the sinusoidal components extracted for this study.

Four low-pass filters with cut-off frequency at 1 k, 1.5 k, 3 k, and 6 kHz were applied to the SM processed AzBio sentences (16 kHz sampling frequency) to create the four BW conditions. In each BW condition, the maximum number of sinusoidal components were limited to 1, 2, and 6, and resynthesized back to sine-wave speech for each of these three sinusoidal component conditions. Figure **1B** illustrates a sentence processed with maximum number of spectral tracks of sinusoidal components was limited to 2 in the frequency band between 0 kHz and 3 kHz, using the low-pass filter with cut-off frequency of 3 kHz. The selected sinusoidal components are represented as red dots on the spectrogram.

Even though we limited the maximum number of sinusoidal components to 1, 2 and 6, for the four BW conditions, the number of sinusoidal components remained in the BW conditions are still dependent on the number of sinusoidal components (after initial selection by SM algorithm) available within the BW in quest.

To identify the number of components selected for 1, 2, and 6 sinusoidal component and 1 k, 1.5 k, 3 k, and 6 kHz low-pass filtering conditions, we performed acoustic analysis. A sample of two AzBio sentences, one spoken by a male speaker and the other spoken by female speaker, were separately processed by the SM algorithm to extract the sinusoidal components at each time frame of the sentence. The extracted sinusoidal components of each sentence were further constrained by each of the BW and maximum number of sinusoidal component conditions. Figure **2** shows the frequency distribution of each sinusoidal components available in each BW condition. The top panel A was plotted with the data obtained from the sentence spoken by the female speaker and the bottom panel B was plotted with data obtained from the sentence spoken by the male speaker.

In each plot, the first sinusoidal component on the x-axis (from left) is the sinusoidal component of the highest magnitude, which also occurs in almost all time frames of the sentence. Magnitude and occurrence in each time frame decreases when the sinusoidal components go to higher order. For instance, the 6th sinusoidal component in the condition to the right of x-axis will have the lowest magnitude among the selected sinusoidal components, and least occurrence in each time frame. The numbers in green boxes on top of each plot indicate the total number of occurrence for each sinusoidal component in the sentence. When the BW decreases from 6 kHz to 1 kHz, the total number of sinusoidal components become less in narrower bandwidth. For instance, it is not possible to limit the maximum number of sinusoidal components to 6 in the BW conditions of 1 kHz, 1.5 kHz, and 3 kHz, with the data obtained from the female speaker (A); and the BW condition of 1000Hz, with the data obtained from the male speaker (B). The sinusoidal components selected by the SM algorithm distributed over a wider frequency

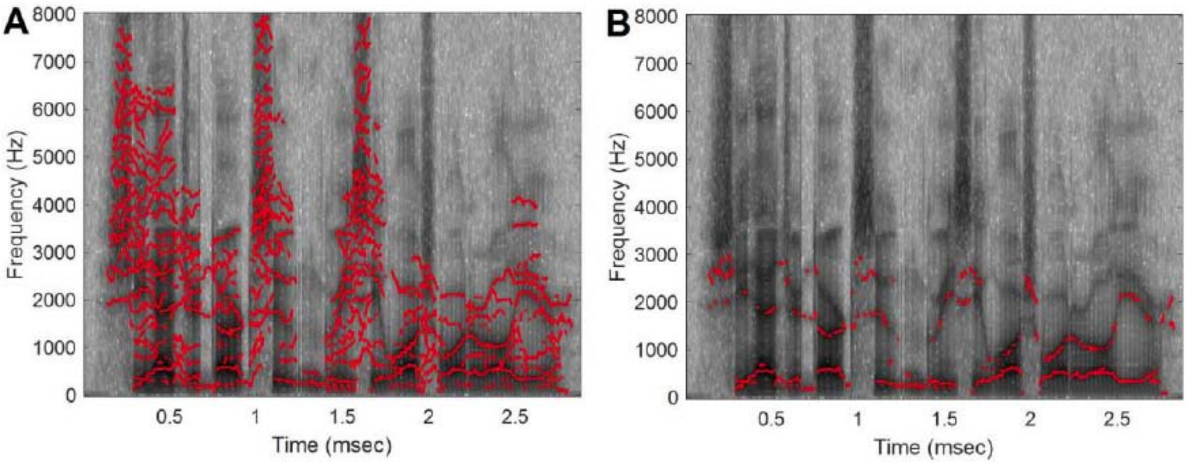

Figure 1: A) The spectrogram showing sinusoidal peaks selected by SM processing (red tracks), **B**) two sinusoidal peaks are selected at each time frame after 3 kHz low pass filtering.

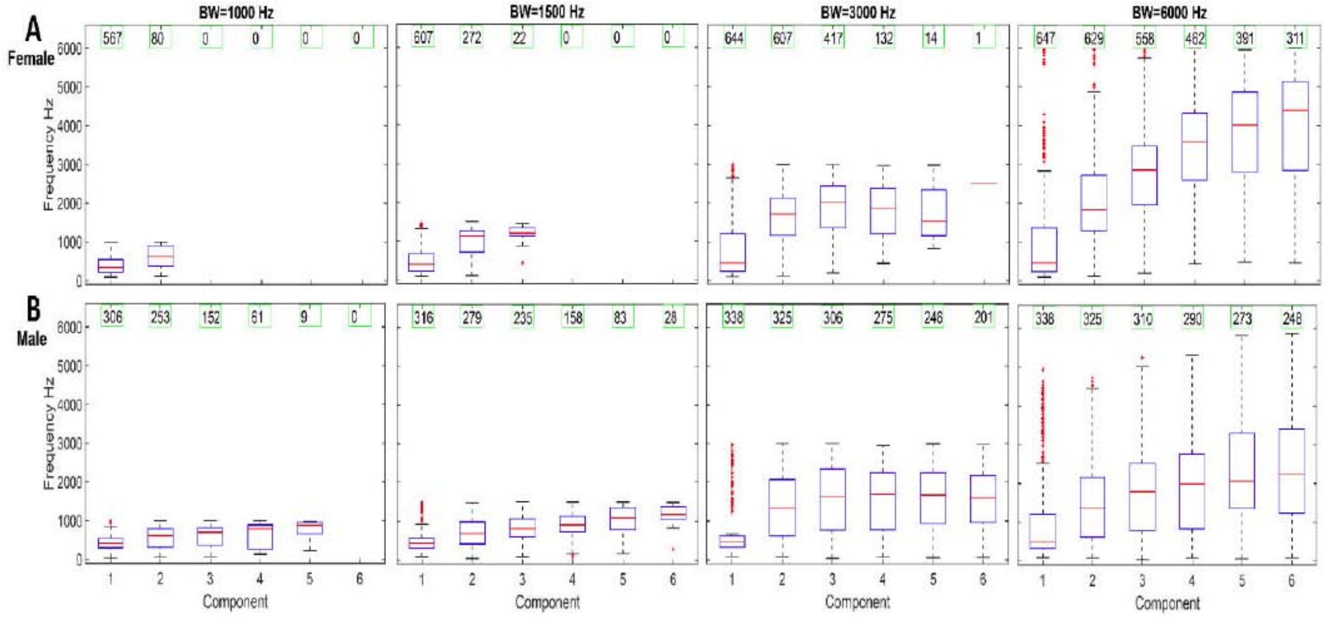

Figure 2: Acoustic analysis of a sample of AzBio sentence processed by SM and filtering conditions. Each box-plot represents the amount of bins and its frequency ranges for selected spectral component from 1 to 6. The plots are displayed for female speaker in the first row and male speaker in the second row, for narrower BW in left columns and broader BW in right columns.

range for the sentence produced by the male talker (B) than the sentence produced by the female talker (A) for all conditions.

Procedures

Speech Perception Test

Testing was conducted when the listener sits in the center of a double-walled sound booth, facing the front speaker placed at 1 m from the listener. 12 lists of sentences were randomly selected from the AzBio database and processed by the SM algorithm for different combination of maximum number of sinusoidal components conditions and BW conditions. The stimuli were presented at 65 dB SPL via the front speaker (0^o) in randomized order. For the two ‘CI’ groups, NH listener with CI simulation (noise vocoded speech) and CI users, one additional list of unprocessed sentences from AzBio database was included in their speech perception test. The listener was asked to repeat the sentence at each time after the test stimulus is presented. For some of the stimuli which are barely intelligible, listeners were encouraged to repeat the words in the sentence as many as possible. Speech perception scores were obtained in percentage based on the number of words correctly identified over the number of words presented.

Speech Quality Rating

At the end of the speech perception test, each subject was asked to rate the perceived sound quality of each sentence on the scale of 1(unnatural) to 10(natural). The greater the number, the better the perceived sound quality of the sentence. Two concatenated sentences, each sentence spoken by a male speaker and one by female speaker selected from the AzBio lists, were processed under the 12 conditions and presented at 65 dB SPL via the front speaker (0^o) in a randomized order. Each listener was asked to perform speech quality rating task twice with stimuli present in two randomized order, and the ratings from two trials were averaged.

EEG Recording and Preprocessing

We recorded the cortical activities of 6 NH subjects listening to CI simulation (noise vocoded speech) of the SM speech and 3 CI subjects listening to the original SM speech, using EEG. The EEG recordings were separately collected for each of the 12 speech stimuli (three maximum number of sinusoidal components conditions and four BW conditions). The stimulus was a sentence from the AzBio database, spoken by a female speaker with a total duration of 2 seconds. The sentence was processed with the 12 conditions that were used in the behavioral tasks.

A 64-channels (actiCAP) electrode cap with scalp electrodes arranged in the modified 10-20 system, was placed on the scalp of the listener, while listening to the stimulus. The impedances of all the electrodes were maintained lower than 10 kΩ. For CI users, the scalp electrodes in the vicinity of CI device (e.g., TP8, P8, P7, TP7) were deactivated. During the recordings, listeners were asked to keep body movements minimal. A silent movie with captions was played while speech stimuli were presented at 65 dB SPL from the frontal speaker. Each sentence was repeatedly presented 150 times with inter-stimulus-interval of 500 ms. The EEG data were recorded with sampling rate of 1 kHz using the BrainVision Recorder.

Further processing was carried out off-line using BrainVision Analyzer 2.1. The data were first re-referenced with two mastoid channels (TP9 and TP10). Then ocular correction with ICA was implemented to eliminate artifacts by eye blinks. 150 epochs from Cz electrode were exported as EEG coefficients (ASCII format) for each listener. Each epoch was filtered between 0.3 and 30 Hz using sixth-order Butterworth filter.

Speech Envelope Extraction

Speech envelope is typically regarded as a low frequency rhythmic fluctuation of speech. The envelope of the stimuli were extracted by computing the magnitude of Hilbert transform. To match the sampling rate of the EEG signal, the speech envelope was resampled to the sampling rate of 1 kHz (sampling rate of EEG signal), and band-pass filtered between 0.3 to 30 Hz using the same sixth-order Butterworth filter used for filtering the EEG signals.

Cross-Correlation between Speech Envelope and EEG Response

To determine whether neural responses follow the speech envelope, the cross-correlation between speech envelope and EEG epochs were calculated for time lags between -200 to 400 msec. For each subject and each test condition, the baseline of the speech envelope and EEG epoch were corrected to zero. For each epoch, the maximum value of cross-correlation was considered as the best match between speech envelope and neural response, and chosen as the maximal cross-correlation value between them. For rest of the manuscript, this value will be referred as the 'maximal correlation'. In each subject and each test condition, we obtained 150 maximal correlation values between 150 EEG epochs and the speech envelope in quest. These maximal correlation values can vary over a wide range. To ensure that we capture the degree of synchrony occurs in most of the epochs, we only considered the data within 80% central confidence interval (120 epochs) and 10% at both tails (30 epochs) were discarded. Figure **3** demonstrate the degree of synchrony by overlaying the speech envelope with different processed EEG signals, showing high (A), moderate (B), and low (C) cross-correlation.

Statistic

Analysis of Variance (ANOVA) was primarily used for our data analysis. In each of our behavioral (speech perception scores and quality ratings) and EEG outcomes, a three way measure ANOVA was used to determine the effects of groups, the number of sinusoidal components, and BWs on speech perception scores and neural maximal correlation respectively for NH and NH with CI simulation groups. To identify the degree of covariance between speech perception scores and maximal correlation, we used Pearson correlation coefficient.

RESULTS

Speech Perception Scores

Mean speech perception scores for three different groups (NH, NH with CI simulation, and CI) are shown in Figure **4**. In general, intelligibility of the sine-wave speech built by small number of sinusoidal components is considerably lower than that of the unprocessed speech. The speech perception scores for unprocessed speech with the NH with CI simulation group and the CI group were found to be 89% and 93%, respectively. Given the maximum number of sinusoidal components limited to 6 with the widest bandwidth of 6 kHz, the speech perception scores for sine-wave speech built under all 12 test conditions used in the study were not higher than 70% correct as expected. However, a monotonic increasing trend in the speech perception scores was observed for the three groups when the maximum number of selected sinusoidal components and the BW increases.

The three-way measures ANOVA examine the effects of groups, the number of sinusoidal components, and BWs on speech perception scores for NH and NH with CI simulation groups. The mean speech perception score for NH group was statistically higher than NH with CI simulation group [F (1, 120) = 178.226, p < .001]. The score was positively associated with the number of sinusoidal components [F (2, 120) = 76.017, p < .001], and BWs [F (3, 120) = 190.185, p < .001]. Post hoc tests using the Bonferroni correction revealed that all pairs of comparisons with number of sinusoidal components and BWs were significantly different from each other (p < .05), except

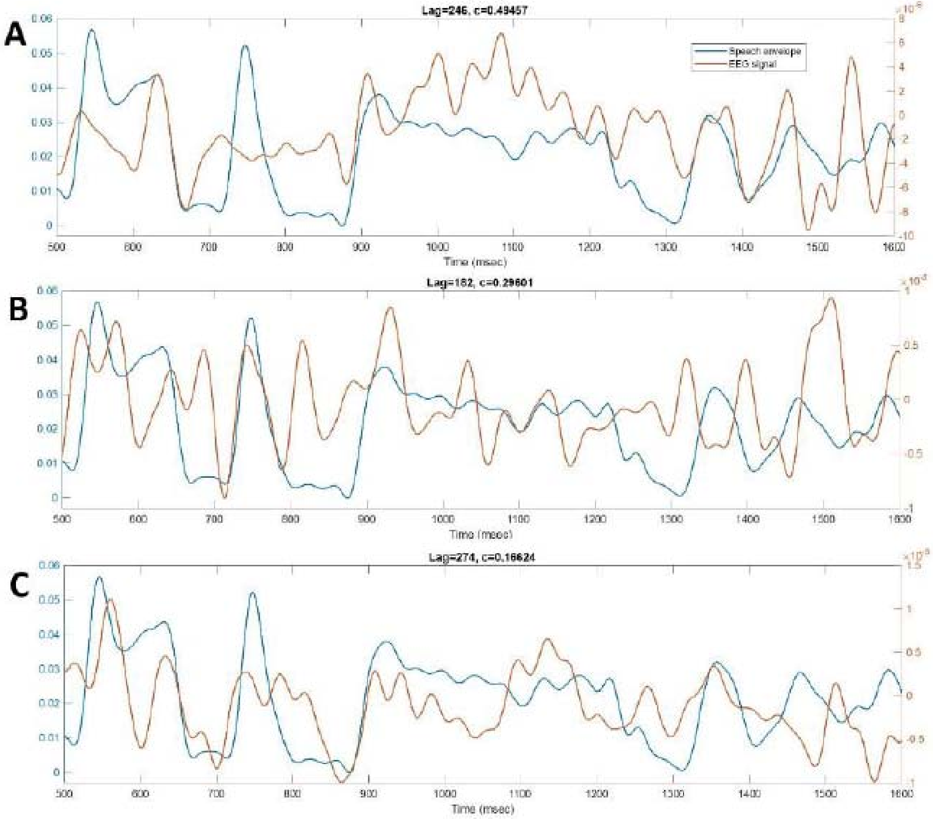

Figure 3: Three sample segments of speech envelope (blue) overlapping with EEG response (red) that represent high (**A**), moderate (**B**), and low (**C**) cross-correlations. The EEG activities were obtained from CI subject 1 in response to the speech processed with 1 sinusoidal component and the low-pass filter at 1 kHz. The cross-correlation was carried out with 2 second of speech envelope and 2.5 second of EEG epoch, but here we capture the middle part of the match-up (500 msec – 1,600 msec) to show clear overlaps. As illustrated, the degree of match-ups (maximal correlations) and corresponding time lags for the different EEG epochs in response to the same stimuli are not identical.

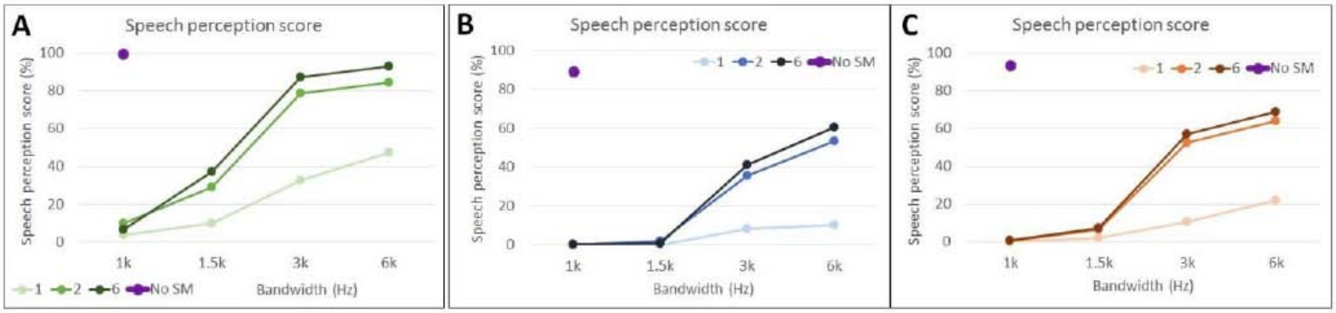

Figure 4: Speech perception scores for three groups: 6 NH listeners who listen to SM speech (**A**, green), 6 NH listeners who listen to SM speech processed with CI simulation (**B**, blue), and 4 CI listeners who listen to SM speech (**C**, red).

for the pair of sinusoidal components 2 vs. 6 (P = .197). Significant interactions for all pairs of the three factors were found with group * number of sinusoidal components [F (2, 120) = 3.228, p < .05], group * BW [F (3, 120) = 13.168, p < .001], and number of sinusoidal components * BW [F (6, 120) = 13.246, p < .001], but when not all three together, group * number of sinusoidal components * BW [F (6, 120) = 1.494, p = .186].

Quality Ratings

Figure **5** shows group mean speech quality ratings for the three different groups. Similar to the trend of the speech perception scores, fewer number of sinusoidal components and narrower BW were associated with poorer sound quality ratings reported. Another three-way measure ANOVA was conducted to examine the effect of the same three factors on sound quality ratings. The two groups (NH and NH with CI

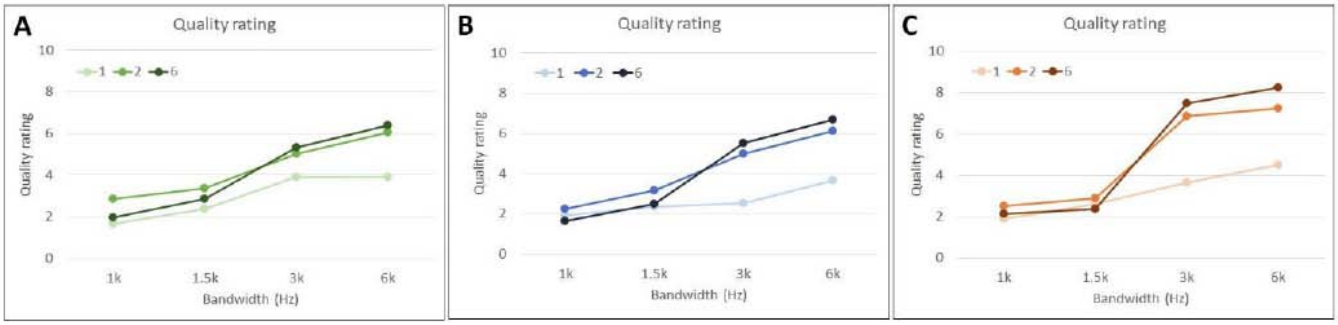

Figure 5: Speech quality ratings for three groups: 6 NH listeners who listen to SM speech (**A**, green), 6 NH listeners who listen to SM speech processed with CI simulation (**B**, blue), and 4 CI listeners who listen to SM speech (**C**, red).

simulation) were not significantly different [F (1, 120) = .007, p = .936], but significant main effects were found for the number of sinusoidal components [F (2, 120) = 11.379, p < .001], and BW [F (3, 120) = 33.115, p < .001]. A Bonferroni test was used for post hoc analysis. It reported that 2-component condition was higher than 1-component condition, and 6-component condition was higher than 1-component condition, (p < .05), but 2-component and 6-component condition was not significantly different (p = 1.000). For BW, all pairs were significantly different from each other (p < .05) with the broader BW, the higher quality, except for the pair 1 vs. 1.5 kHz (p = .573). No interaction was found between the three factors.

EEG Result

Figure **6** shows the regression of maximal correlations between EEG response and speech envelope (indicated by red asterisks) as a function of number of sinusoidal components actually remained in the four BWs (1 k, 1.5 k, 3 k and 6 kHz) in CI users and NH listeners with CI simulation. R values and corresponding p-values of the regression are shown for each condition. Speech perception scores for corresponding conditions are plotted as blue bars for comparisons. Different length of regression lines were plotted according to number of sinusoidal components remained in each BW condition, based on the acoustic analysis outcome on the sine-wave sentences. In both CI and NH with CI simulation group, maximal correlation increases with the broadening of the BW, which is consistent with their behavioral results. Likewise, the maximal correlation between EEG signal and speech envelope increases when the number of sinusoidal components increases; with an 'unclear' exception in 1 kHz BW condition for CI group.

Since different number of sinusoidal components remained in each BW conditions as shown in Table **1**, only the same sample size in conditions, where number of components are 1 and 2, with our current subject population, allowed some statistical analyses to be performed.

The three way ANOVA was conducted with maximal correlations as the dependent variable, and groups (CI users and NH with CI simulated listeners), number of sinusoidal components (1 and 2), and BWs (1 k, 1.5 k, 3 k, and 6 k) as independent variables. There was a significant main effect on BWs [F (3, 7386) = 1103.84, p < .001], but no significant main effect on groups [F (1, 7386) = .04, p = .842], and number of sinusoidal components [F (1, 7386) = 1.02, p = .313]. Interaction effect was found with groups vs. number of sinusoidal components [F (1, 7386) = 10.98, p < .001], and groups vs. BWs [F (3, 7386) = 237.69, p < .001], but not with number of sinusoidal components vs. BWs [F (3, 7386) = 1.11, p = .344]. A two-way ANOVA was performed to further separate the effect of group with the factors of number of sinusoidal components and BWs. For CI group, 2 component was significantly higher than 1 component in terms of maximal correlation [F (1, 2841) = 7.26, p < .001]. There was also main effect of BWs [F (3, 2841) = 123.04, p < .001], with post-hoc pairwise comparison leveling that higher BW resulted in

Table 1: Maximum number of sinusoidal components limited within the bandwidths (2nd row) and maximum number of sinusoidal components remained in the bandwidth after the component limitation (3rd row)

Bandwidth conditions	1 kHz			1.5 kHz			3 kHz			6 kHz		
Maximum number of sinusoidal components allowed	1	2	6	1	2	6	1	2	6	1	2	6
Maximum number of sinusoidal components remained	1	2	2	1	2	3	1	2	5	1	2	6

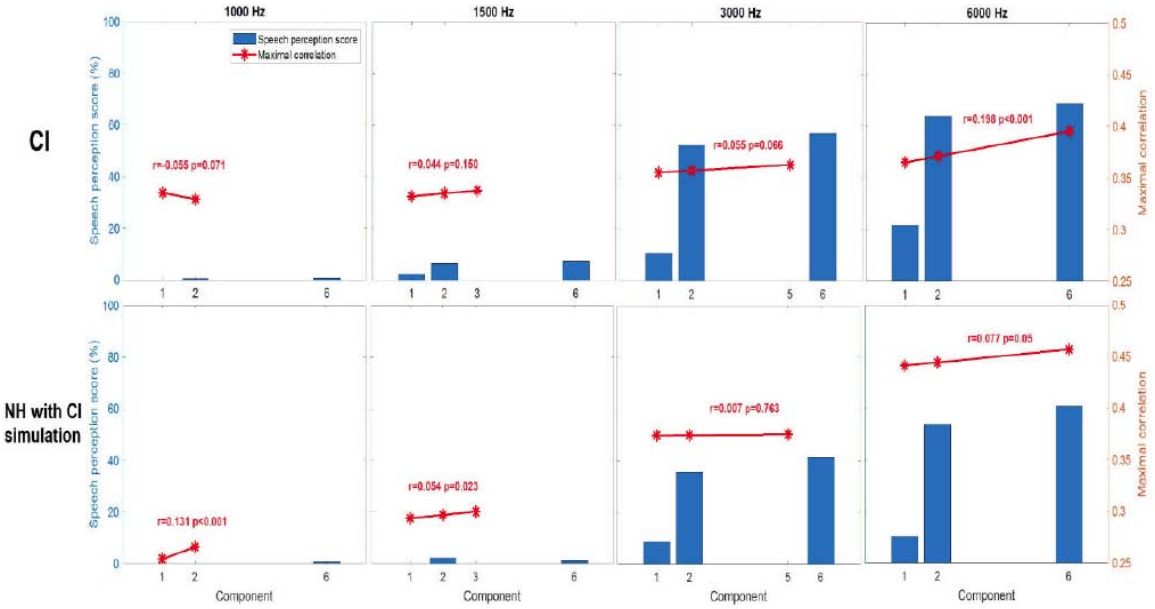

Figure 6: Cross-procedure comparisons between behavioral and physiological results. Speech perception scores (blue bar graph) and regression lines of maximal correlations (red regression line) are plotted as a function of sinusoidal components. Right Y-axis denotes speech perception scores (%) and Left Y-axis denotes maximal correlation between speech envelope and cortical response. The plots are displayed for CI group in the first row and NH group in the second row, for narrower BW in left columns and broader BW in right columns.

significantly higher maximal correlation except for the pairs of 1 kHz and 1.5 kHz (P < .05). There was no interaction effect between the two variables [F (3, 2841) = .42, p = .741]. For NH group listening to CI simulated speech, significant increase in maximal correlation was found with increase in BWs [F (3, 4542) = 1595.84, p < .001], and post-hoc comparisons revealed that all pairs were significantly different. No main effect was found for number of sinusoidal components [F (1, 4542) = 3.74, p = .053]. There was an interaction effect between number of sinusoidal components and BWs [F (3, 4542) =3.46, p = .0157].

To determine the relationship between the speech intelligibility and cortical tracking in CI users, the speech perception scores and corresponding maximal correlations for all subjects (3 CI users and 6 CI simulated NH listeners) in test conditions of 1 and 2 components and four BWs are plotted in Figure **7**. The EEG and speech scores were normalized using the z-score due to large scale difference between maximal correlations and speech perception scores (raw speech perception scores ranged between 0 and 90 and raw maximal correlation ranged between 0.45 and 0.76). Pearson correlation indicates that speech perception score is significantly correlated with maximal correlation between EEG signal and speech envelope both for CI and CI simulated NH subjects (r = .72, p < .001 for CI, and r = .63, p < .001 for CI simulated NH subjects).

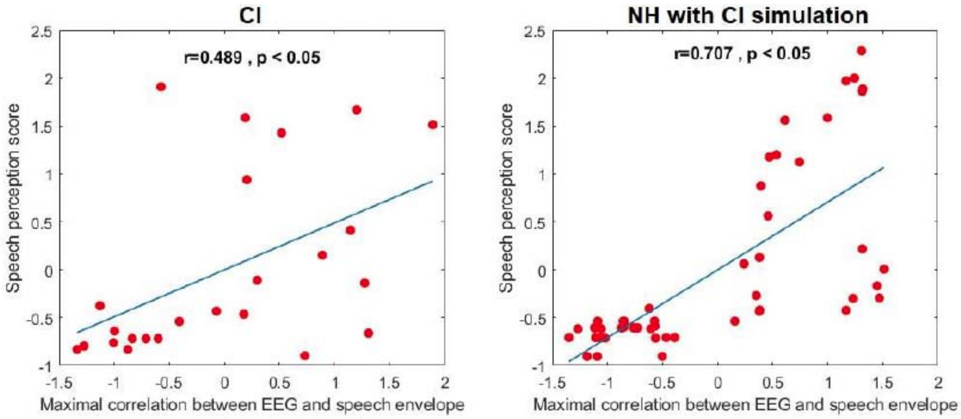

Figure 7: Relation between behavioral and physiological results. Data for two different scales was z-score normalized.

DISCUSSION

Perceptual Effect of SM

Comparatively, the NH control group showed the best speech perception performance among the three groups; and the CI group has a better performance than the NH with CI simulation group. This coincides with our preliminary findings [1] that intelligibility of sine-wave speech is better for NH listeners than CI users. Perceptual difference between the two groups was also reported in noise and varying loudness conditions [25]. Observed higher speech perception scores for CI group over CI simulated group may be attributed to the degree of familiarity with the experimental stimuli. Unlike CI users who have been exposed to spectrally degraded speech in their everyday life, NH group probably was not able to perceive such odd speech sound proficiently. CI stimulation using noise vocoded speech may not fully mimic the CI mediated hearing to NH listener.

Intelligibility of sine-wave speech is expected to be lower than that of the natural speech as sine-wave speech is only resynthesized from the small number of spectral components. The limited number of time-varying sinusoids can only convey frequency and amplitude information in speech, loosing other critical features such as harmonics, broad band formant transitions, and fundamental frequencies [4,26]. Narrowing BW in the study not only limit the number of sinusoidal components for sine-wave speech re-synthesis, but also examine whether the small number of sinusoidal components in low frequencies is able to deliver all relevant speech landmarks in broad spectrum. In a wider bandwidth of 6 kHz and 6 sinusoidal components, we were able to resynthesize a more intelligible sine-wave speech with NH group scoring 90% correct and CI group scoring 60% in their speech perception test. Quatieri and McAulay [5] also claimed that speech built with nearly 80 sinusoidal components is indistinguishable from the original speech. SM is able to retain most of the intelligibility and quality of speech that is enough for daily communication, with limited amount of speech information. The speech cues carried in these limited number of time varying sinusoidal components could possibly be the essential cues needed to retain in a channel limited hearing devices like cochlear implant and hearing aid.

EEG Results

In this study we used speech envelope tracking property of cortical responses to investigate the underlying neural mechanism of CI user and NH listener in encoding sine-wave speech and noise vocoded sine-wave speech. Since both CI speech processor and noise vocoder encodes only envelope information of speech, it is logical to choose cortical entrainment to speech envelope fluctuations as a measure to relate the perceptual entities like intelligibility and quality of speech convey by the encoded information.

As previously described, maximal cross-correlation is used as an index to quantify the tracking relationship between speech envelope and corresponding neural activity. Our result indicates that the degree of neural entrainment to speech envelope reflects the intelligibility and sound quality degradation associated with the distorted speech envelop. This finding is in alignment with the outcome of previous works showing that cortical entrainment to speech envelopes is positively related to speech intelligibility [14,27] even when spectrally degraded speech was presented [21,28]. In our study, the maximal correlation computed between the EEG signals and speech envelope was higher in value with the broader BW. The sinusoidal components selected by the SM algorithm are usually well distributed across the spectrum. More sinusoidal components will be retained in the broader BW condition to build a sine-wave speech that is close in intelligibility to the original speech, which also leads to higher maximal correlation. Our result also suggests a balanced mix of high and low frequency sinusoidal components is necessary to build a more intelligible SM speech of higher perceived sound quality.

With CI simulation of sine-wave speech built with 1 and 2 sinusoidal components, the difference between the two corresponding cortical entrainment patterns obtained from this group of NH listeners was found not to be statistically significant. Both their speech perception score and perceived quality rating do show a significant monotonic increase when the number of sinusoidal components increases in this group. However, a different observation was found with the CI group: their maximal correlations with EEG were found to be significantly associated with sinusoidal components. The insignificant effect of sinusoidal components in our statistics may be attributed to the narrow band conditions, 1 kHz and 1.5 kHz, which exhibit extremely low speech perception scores. These unintelligible speeches may not be enough to trigger robust neural tracking responses associated with spectral components.

Pearson correlational analysis showed that our behavioral speech perception outcomes and maximal neural correlations are highly correlated regardless of stimulus conditions (Figure **7**). This is in agreement with other studies [21,29] in which neural tracking is still observable in response to spectro-temporally degraded speech. Relatively higher correlation between maximal correlations and speech perception scores was found in the NH listeners with CI simulated speech (r = .707) than the CI users (r =.489). In our behavioral study, speech recognition test required subject to actively engaged in the task; even guessing the words or sentences to score. However, our EEG recordings were obtained when subject was listening to the stimuli passively. Subject's active attention to stimulus will be included in the following phase of investigation. Another aspect to consider is the heterogeneity in EEG recordings obtained in acoustic and electrical listening and artifacts arises from CI device. In addition, CI stimulation modeled using the noise vocoder may not depict the difference in mechanism between electrical and acoustic listening. Javel [30] showed that electrical stimulation by CIs elicits phase locked responses of auditory nerves at higher frequencies compared to acoustic stimulation. The above variables will be included in the following phase of our study for further investigation.

CONCLUSION

In present study, behavioral speech intelligibility and quality ratings for band-limited sine-wave speech were assessed by 6 NH control subjects and 6 NH subjects with CI-simulation and 4 CI users. An 8 channel noise vocoder was used as a CI simulation for NH with CI simulation group. EEGs were also measured to investigate how cortical activities synchronize to the sine-wave speech re-synthesized using the different number of sinusoidal components and band-widths. The outcome of this study shows that sine-wave speech built with a small number of sinusoidal component can deliver a reasonably good intelligibility and perceived sound quality for daily communication. Similar to NH listeners, cortical entrainment to the speech envelope is found in CI users when they are listening sine-wave speech. Considerably high correlations were found between speech perception scores and maximal correlations for the CI group and the NH with CI simulation group. More CI users listening to sine-wave speech built with more sinusoidal components will be also explored in future study to derive more insights in retaining the essential cues for assistive hearing devices with limited channel processing.

ACKNOWLEDGEMENTS

This study was supported in whole by from the Texas States Research Grants.

REFERENCES

[1] Lee S, Akbarzadeh S, Singh S, Tan C. A speech processing strategy based on sinusoidal speech model for cochlear implant users 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Honolulu, HI, USA: IEEE 2018; pp. 393-7. https://doi.org/10.23919/APSIPA.2018.8659620

[2] Hillenbrand JM, Clark MJ, Baer CA. Perception of sinewave vowels. J Acoust Soc Am 2011; 129(6): 3991-4000. https://doi.org/10.1121/1.3573980

[3] Carrell TD, Opie JM. The effect of amplitude comodulation on auditory object formation in sentence perception. Percept Psychophys 1992; 52(4): 437-45. https://doi.org/10.3758/BF03206703

[4] Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science (80-) 1981; 212(4497): 947-9. https://doi.org/10.1126/science.7233191

[5] McAulay RJ, Quatieri TF. Speech analysis/Synthesis based on a sinusoidal representation. IEEE Trans Acoust 1986; 34(4): 744-54. https://doi.org/10.1109/TASSP.1986.1164910

[6] Kates JM. Speech enhancement based on a constrained sinusoidal model. J Speech, Lang Hear Res 1994; 37(2): 449-64. https://doi.org/10.1044/jshr.3702.449

[7] Timms O. Speech processing strategies based on the sinusoidal speech model for the profoundly hearing impaired. Dr Diss 2003.

[8] Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 2012; 107(1): 78-89. https://doi.org/10.1152/jn.00297.2011

[9] Di Liberto GM, O'sullivan JA, Lalor EC. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 2015; 25(19): 2457-65. https://doi.org/10.1016/j.cub.2015.08.030

[10] Khalighinejad B, Cruzatto da Silva G, Mesgarani N. Dynamic encoding of acoustic features in neural responses to continuous speech. J Neurosci 2017; 37(8): 2176-85. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

[11] Riecke L, Formisano E, Sorger B, Başkent D, Gaudrain E. Neural entrainment to speech modulates speech intelligibility. Curr Biol 2018; 28(2): 161-169.e5. https://doi.org/10.1016/j.cub.2017.11.033

[12] Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 2007; 54(6): 1001-10. https://doi.org/10.1016/j.neuron.2007.06.004

[13] Aiken SJ, Picton TW. Human cortical responses to the speech envelope. Ear Hear 2008; 29(2): 139-57. https://doi.org/10.1097/AUD.0b013e31816453dc

[14] Ding N, Simon JZ. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 2013; 33(13): 5728-35. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

[15] Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. J Neurophysiol 2013; 109(12): 3082-93. https://doi.org/10.1152/jn.01026.2012

[16] Kong YY, Mullangi A, Ding N. Differential modulation of auditory responses to attended and unattended speech in different listening conditions. Hear Res 2014; 316: 73-81. https://doi.org/10.1016/j.heares.2014.07.009

[17] O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, *et al*. Attentional selection in a cocktail party environment can be decoded from single-Trial EEG. cereb cortex 2015; 25(7): 1697-706. https://doi.org/10.1093/cercor/bht355

[18] Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 2015; 19(1): 158-64. https://doi.org/10.1038/nn.4186

[19] Ding N, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front Hum Neurosci 2017; 11: 1-9. https://doi.org/10.3389/fnhum.2017.00481

[20] Kösem A, van Wassenhove V. Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Lang Cogn Neurosci 2017; 32(5): 536-44. https://doi.org/10.1080/23273798.2016.1238495

[21] Kong YY, Somarowthu A, Ding N. Effects of dpectral fegradation on sttentional modulation of vortical suditory responses to continuous speech. JARO - J Assoc Res Otolaryngol 2015; 16(6): 783-96. https://doi.org/10.1007/s10162-015-0540-x

[22] Verschueren E, Somers B, Francart T. Neural envelope tracking as a measure of speech understanding in cochlear implant users. Hear Res 2019; 373: 23-31. https://doi.org/10.1016/j.heares.2018.12.004

[23] Kumagai Y, Matsui R, Tanaka T. Music familiarity affects EEG entrainment when little attention is paid. Front Hum Neurosci 2018; 12: 1-11. https://doi.org/10.3389/fnhum.2018.00444

[24] Spahr AJ, Dorman MF, Litvak LM, Van Wie S, Gifford RH, Loizou PC, *et al*. Development and validation of the AzBio sentence lists. Ear Hear 2012; 33(1): 112-7. https://doi.org/10.1097/AUD.0b013e31822c2549

[25] Akbarzadeh S, Lee S, Chen F, Tan C. The effect of perceived sound quality of speech in noisy speech perception by normal hearing and hearing impaired listeners 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin, Germany, pp. 3119-22. https://doi.org/10.1109/EMBC.2019.8857952

[26] Goh WD, Pisoni DB, Kirk KI, Remez RE. Audio-visual perception of sinewave speech in an adult cochlear implant user: A case study. Ear Hear 2001; 22(5): 412-9. https://doi.org/10.1097/00003446-200110000-00005

[27] Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T. Speech intelligibility rredicted from neural entrainment of the speech envelope. JARO - J Assoc Res Otolaryngol 2018; 19(2): 181-91. https://doi.org/10.1007/s10162-018-0654-z

[28] Ding N, Chatterjee M, Simon JZ. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 2014; 88: 41-6. https://doi.org/10.1016/j.neuroimage.2013.10.054

[29] Müller JA, Wendt D, Kollmeier B, Debener S, Brand T. Effect of speech rate on neural tracking of speech. Front Psychol 2019; 10: 1-15. https://doi.org/10.3389/fpsyg.2019.00449

[30] Javel E. Acoustic and electrical encoding of temporal information. Cochlear Implant. New York: Springer 1990; pp. 247-95. https://doi.org/10.1007/978-1-4612-3256-8_17

Received on 03-05-2020 Accepted on 16-05-2020 Published on 10-06-2020

DOI: https://doi.org/10.12970/2311-1917.2020.08.06