HHS Public Access

Author manuscript

Brain Lang. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

Brain Lang. 2021 February ; 213: 104891. doi:10.1016/j.bandl.2020.104891.

Neural tracking of the speech envelope is differentially modulated by attention and language experience

Rachel Reetzke[1,2], **G. Nike Gnanateja**[3], **Bharath Chandrasekaran**[3]

[1]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine

[2]Center for Autism and Related Disorders, Kennedy Krieger Institute

[3]Department of Communication Sciences and Disorders, University of Pittsburgh

Abstract

The ability to selectively attend to an incoming speech signal amid competing sounds is a significant challenge, especially for listeners trying to comprehend non-native speech. Attention is critical to direct neural processing resources to the most essential information. Here, neural tracking to the speech envelope of an English story narrative and cortical auditory evoked potentials (CAEPs) to non-speech stimuli were simultaneously assayed in native and non-native listeners of English. Although native listeners exhibited higher narrative comprehension accuracy, non-native listeners exhibited enhanced neural tracking of the speech envelope and heightened CAEP magnitudes. These results support an emerging view that although attention to a target speech signal enhances neural tracking of the speech envelope, this mechanism itself may not confer speech comprehension advantages. Our findings further suggest that non-native listeners may engage neural attentional processes that enhance low-level acoustic features, regardless if the target signal contains speech or non-speech information.

Keywords

neural speech tracking; amplitude envelope; attention; language experience; speech comprehension

Correspondence to: Bharath Chandrasekaran, Department of Communication Sciences and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, 6036 Forbes Tower, Pittsburgh, PA 15260, b.chandra@pitt.edu.

AUTHOR CONTRIBUTIONS

Conceptualization, B.C. and R.R.; Methodology, R.R. and B.C.; Data Curation, R.R. and G.N.G.; Formal Analysis, R.R. and G.N.G.; Investigation, R.R. and B.C.; Resources, B.C.; Writing – Original Draft, R.R., G.N.G., and B.C.; Writing – Review & Editing, R.R., G.N.G., and B.C.; Visualization, R.R., and G.N.G.; Project Administration, R.R. and B.C.; Funding Acquisition, B.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

1. INTRODUCTION

During everyday communication, several factors compromise speech comprehension, ranging from distracting sounds in the environment to limited language proficiency of the listener (for a review see, Mattys et al., 2012). To attend to a specific speech signal amid competing sound sources, the auditory system must first extract key acoustic features from the incoming signals to segregate the target speech signal from the distractor (Bregman, 1994). Accurate speech comprehension further depends on the listener's knowledge of the language of the incoming speech signal. Limited understanding of the linguistic rules governing the incoming speech signal can result in the inaccurate mapping of incoming sensory information onto linguistic units, ultimately resulting in misinterpretation of the speech signal (Ahissar et al., 2001; McClelland & Elman, 1986). The ability to selectively attend to an incoming speech signal amid competing sounds is a significant sensory and cognitive challenge, especially for listeners trying to comprehend non-native speech (for reviews see, Lecumberri et al., 2010; Scharenborg & van Os, 2019).

Converging evidence suggests that non-native listeners exhibit lower performance on speech comprehension tasks under various listening conditions, compared to native listeners (Bidelman & Dexter, 2015; Brouwer et al., 2012; Cooke et al., 2008; Cutler et al., 2004; Mayo et al., 1997; Rogers et al., 2006). This body of literature outlines at least two possible explanations for poorer speech comprehension in non-native listeners in challenging listening environments. Reduced speech comprehension in non-native listeners may be due to imprecise encoding of non-native acoustic cues at the early stages of sensory processing. For example, non-native listeners may be less sensitive to the language-specific acoustic cues of their non-native language (Flege & Hillenbrand, 1986; Mayo et al., 1997). This, in turn, may cause difficulty segregating the speech signal from the distracting sound sources early in sensory processing. Poorer speech comprehension could also be the result of differential allocation of neural attentional resources in non-native listeners, relative to native listeners. For example, behavioral evidence suggests that when segmenting spoken English under cognitively demanding tasks, non-native listeners may exploit low-level acoustic cues rather than higher-level lexical-semantic information due to limited word knowledge (Mattys et al., 2010). Other neural evidence suggests that while native listeners allocate more attention to the onset of the words, proficient non-native listeners show increased attention to the entire duration of spoken sentences (Astheimer et al., 2016). Taken together, these studies suggest that non-native listeners may exhibit lower performance on speech comprehension tasks due to differential allocation of neural attentional resources relative to monolingual, native listeners.

Slow fluctuations in the amplitude of the incoming speech signal over time, or the speech envelope, is a critical acoustic cue that has been found to contribute to speech comprehension (Fu et al., 1998; Rosen, 1992; Shannon et al., 1995). Converging evidence from electroencephalography (EEG; Hambrook & Tata, 2014; Kerlin et al., 2010; O'Sullivan et al., 2014), magnetoencephalography (MEG; Ding & Simon, 2011, 2012), and intracranial electrocorticography (ECoG; Golumbic et al., 2013) have indicated that low-frequency cortical activity reliably tracks the speech envelope. Neural tracking of the speech envelope measured via EEG, MEG, or ECoG can be achieved by modeling the transformation of the

speech envelope in continuous natural speech to neurophysiological responses (Crosse et al., 2016; Xie et al., 2019). In challenging listening conditions such as those with more than one speaker, or background noise, neural tracking of the attended speech envelope is enhanced relative to the unattended auditory stream (Ding & Simon, 2012; Fuglsang et al., 2017; Golumbic et al., 2013; Jess R Kerlin et al., 2010; O'sullivan et al., 2014). Enhanced neural tracking to the attended speech envelope is posited to contribute to speech comprehension in challenging listening environments in two important ways. First, neural tracking of the attended speech envelope has been a hypothesized mechanism by which the attended speech signal is separated from the listening background (Ding & Simon, 2012; Golumbic et al., 2013), as the speech envelope provides important acoustic cues for grouping the acoustic features belonging to the target sound signal (Shamma et al., 2011). Second, neural tracking of the speech envelope corresponds to the rate at which syllables (4–8 Hz) and words/phrases (1–4 Hz) occur, and in turn has been posited to be an important acoustic cue for the segmentation of continuous, dynamic speech signals into meaningful units needed for later stages of speech comprehension (Ding & Simon, 2014; Giraud & Poeppel, 2012). Although tracking of attended speech is often reflected by enhanced neural responses to attended relative to distractor auditory signals, evidence also indicates that that the distractor signal is processed, at least to some extent or form, at early stages of auditory processing (Golumbic et al., 2013).

Based on the existing and emerging literature on the neural tracking of the speech envelope (Ding & Simon, 2014), two prominent hypotheses have been put forward regarding the mechanisms underlying neural tracking of the speech envelope. Per the domain-general auditory encoding hypothesis, neural tracking of the speech envelope is primarily driven by general auditory mechanisms that can also be evidenced in animal models (Doelling et al., 2014; Joris et al., 2004; Steinschneider et al., 2013). An alternate hypothesis (interactive processing hypothesis) suggests that the neural tracking of the speech envelope reflects dynamic interactions between the processing of low-level acoustic cues and higher-level linguistic information (Zou et al., 2019). Evidence for this hypothesis comes from studies demonstrating that neural tracking of the speech envelope is modulated by selective attention, language experience, as well as the predictability of the incoming speech information (Ding & Simon, 2012; Golumbic et al., 2013; Zou et al., 2019). This literature indicates that higher-level cognitive-linguistic factors can alter the neural encoding of low-level acoustic cues. Of particular interest is the effect of language proficiency on neural tracking of the envelope, as this would help in understanding the extent to which language processes interact with the encoding of the low-level acoustic cues.

Recent studies evaluating the effect of language experience on neural tracking have shown increased neural tracking of both non-native and completely unfamiliar speech envelopes (Song & Iverson, 2018; Zou et al., 2019). In contrast to the literature on native neural speech tracking, enhanced neural tracking in non-native and naïve listeners has been associated with *poorer* behavioral speech perception of non-native speech (Song & Iverson, 2018; Zou et al., 2019). The enhanced neural tracking in non-native listeners has been interpreted as increased listening effort and over-reliance on the acoustic cues in speech due to the inability to use higher linguistic structures to perceive speech (Song & Iverson, 2018). In these previous studies have evaluated neural speech tracking using speech presented in competing speech

paradigms, where listeners are required to not only attend to the target speech signal but also expend additional cognitive resources to ignore the linguistically meaningful distractor (Shinn-Cunningham, 2008). In turn, it is unclear if differences due to language experience are driven by increased neural tracking of the attended speech signal or due to difficulty suppressing meaningful information in the competing speech signal.

In the current study, neural responses were recorded from native and non-native listeners of English as they listened to an English narrative simultaneously presented with a harmonic tone sequence. Neural tracking of the speech envelope of the narrative (speech stimuli) and cortical auditory evoked potentials (CAEPs) to a harmonic tone sequence (non-speech stimuli) were measured across two conditions: (1) an attend speech condition, where participants were instructed to pay attention to the story narrative and ignore the harmonic tone sequence; and an (2) attend tone condition, where participants were instructed to attend to the harmonic tone sequence and ignore the story narrative (Figure 1). This experimental paradigm facilitated the interpretation of neural tracking of the attended speech envelope, without the confound of additional linguistic interference from a speech masker. We hypothesized that non-native listeners would exhibit poorer speech comprehension relative to native listeners, based on decades of non-native vs. native listener behavioral speech perception, recognition, and comprehension in noise literature (Cooke et al., 2008; Lecumberri et al., 2010; Mayo et al., 1997). We hypothesized that if poorer speech comprehension in non-native listeners is due to imprecise encoding of non-native acoustic cues at early stages of sensory processing, then non-native listeners should exhibit poorer neural tracking of the speech envelope relative to native listeners. Alternatively, if neural tracking of speech envelope reflects differential allocation of attentional resources in non-native listeners relative to native listeners, then non-native listeners may exhibit *enhanced* neural tracking of the *attended* speech envelope compared to native listeners. Finally, if neural tracking of the speech envelope reflects a general greater reliance on low-level acoustic cues during speech processing in non-native listeners, then non-native listeners should exhibit enhanced neural tracking of the speech envelope, irrespective of condition (attend speech vs. attend tone). The CAEPs provided an estimate of the effect of attention and language experience on the cortical response to a tone sequence that is relatively neutral to both the listener groups.

2 . MATERIAL AND METHODS

Written consent was obtained from all participants before study participation. The Institutional Review Board at the University of Texas at Austin approved all materials and procedures. All procedures were carried out following the approved guidelines. Each participant received monetary compensation for their participation in this study.

2.1 Participants

Thirty young adults were recruited from the University of Texas at Austin student body to participate in the current investigation. Fifteen adult native speakers of English (9 females; $M = 22.53$ years, $SD = 3.66$ years) and fifteen adult non-native speakers of English (7 females; $M = 24.07$ years, $SD = 2.62$ years). One non-native participant's EEG data could

not be used in the current investigation due to excess noise in the EEG data. Therefore, this participant was excluded from the analyses that follow. The decision about excess noise was made based on the visually observable and pervasively present high amplitude spikes in the EEG data.

Each participant completed a music and language background questionnaire (Li et al., 2014). All participants included in the study were: (1) current students of the University of Texas at Austin; (2) right-handed; (3) and had no previous self-reported history or diagnosis of speech, language, or neurodevelopmental disorder. All participants had air and bone conduction thresholds < 20 dB HL at octave frequencies from 250 to 8,000 Hz. Hearing thresholds were confirmed using an Interacoustics Equinox 2.0 PC-Based Audiometer. Past evidence has shown that music training influences speech processing (Bidelman et al., 2011; Wong et al., 2007); therefore, further inclusion criteria consisted of no history of formal music training or no significant music experience, < 6 years of experience (native: $M = 2.07$ years, $SD = 2.37$ years; non-native: $M = 0.86$ years, $SD = 1.83$ years; $F_{1, 27} = 2.33$, $p = .138$, $\eta_p^2 = 0.08$). Native and non-native participants were closely matched in age ($F_{1, 27} = 1.67$, $p = .207$, $\eta_p^2 = 0.06$), sex (χ^2 (1, $N = 29$) = 0.02, $p = 0.867$), and non-verbal intelligence (Native: $M = 117.27$, $SD = 8.72$; Non-native: $M = 121.65$, $SD = 8.94$; $F_{1, 27} = 1.79$, $p = .193$, $\eta_p^2 = 0.06$), as measured by the Kaufman Brief Intelligence Test-Second Edition, KBIT-2 (Kaufman & Kaufman, 2004), matrices subtest (normal intelligence: $M = 100$, $SD = 15$)

Native participants were all native speakers of American English and reported no significant experience (< 1 year) learning or speaking a second language. Non-native participants were identified as late learners of English, and sequential Mandarin-English bilinguals (Newman et al., 2012). All non-native participants were: (1) born and raised in mainland China; (2) spoke Mandarin Chinese as their native language; (3) did not begin learning English formally until after the age of 5 years (range = 6–16 years, $M = 9.86$ years, $SD = 2.68$ years); (4) and lived in the United States for no more than six years (range = 1– 6 years, $M = 2.33$ years, $SD = 1.90$ years).

2.2 English Language Proficiency Assessment

To confirm differences in language proficiency between listener groups, English language proficiency for native and non-native participants was measured with the Test of Adolescent and Adult Language-Fourth Edition (TOAL-4; Hammill et al., 2007). The TOAL-4 is a standardized assessment of language ability that does not typically show ceiling effects among native English speaking adolescents and adults and has proved to be a useful metric of English language proficiency in past language processing based EEG studies (Newman et al., 2012; Pakulak & Neville, 2010; Weber-Fox et al., 2003), particularly in comparing native and non-native, late learners of English (Newman et al., 2012). This assessment tool consists of six subtests (word opposites, word derivations, spoken analogies, word similarities, sentence combining, and orthographic usage). From these six subtests, three composite scores can be derived: spoken language, which reflects oral English language competence; written language, which reflects written English language competence; and general language, which represents overall English language competence (Hammill et al.,

2007). These composite scores are reported as ability scores (standard score with a mean of 100 and a standard deviation of 15).

Each of the three TOAL-4 composite scores were significantly higher for native English-speaking participants relative to the non-native English participants. Across all three TOAL-4 composite scores native participants had higher English language proficiency relative to the non-native participants (spoken: $F_{1, 27} = 110.62$, $p < .001$, $\eta_p^2 = 0.80$; written: $F_{1, 27} = 81.92$, $p < .001$, $\eta_p^2 = 0.75$; general: $F_{1, 27} = 123.18$, $p < .001$, $\eta_p^2 = 0.82$). These results confirmed that the two groups included in the current investigation significantly differed in English language proficiency.

2.3 Stimuli

2.3.1 Speech Stimuli—Speech stimuli were selected from a narration of the novel *Alice's Adventures in Wonderland* (Chapters 1–7, http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-5). The novel was read in American English by a male speaker and sampled at a frequency of 22,500 Hz. The recorded chapters were divided into 60 unique segments, each ~60 s in duration after long speaker pauses (> 500 ms) were shortened to 500 ms. The mean modulation spectrum of the audio segments is shown in Supplementary Figure 1. The modulation spectrum shows a peak corresponding to a modulation frequency of ~6 Hz, which is analogous to the syllable rate (Ding et al., 2017).

Because *Alice's Adventures in Wonderland* is a popular and well-known novel, participants completed a questionnaire before participating in this experiment, indicating if they had: (1) read *Alice's Adventures in Wonderland* or (2) watched any movie adaptation of the novel. The number of participants who had read the book did not significantly differ between native and non-native groups (χ^2 (1, N = 29) = 1.28, $p = 0.257$). Within-language group contrasts examining the extent to which reading or not reading the book predicted story comprehension accuracy, indicated no significant group differences (native participants: $\beta = -0.054$, $SE = 0.060$, $z = -0.900$, $p = 0.368$; non-native participants: $\beta = -0.008$, $SE = 0.087$, $z = -0.096$, $p = 0.924$). The number of participants who watched a movie adaptation of the book also did not significantly differ between native and non-native groups (χ^2 (1, N = 29) = 0.84, $p = 0.360$). Within-language group contrasts examining the extent to which watching or not watching a movie adaptation of the book predicted story comprehension accuracy, indicated no significant group differences (native participants: $\beta = 0.031$, $SE = 0.063$, $z = 0.491$, $p = 0.623$; non-native participants: $\beta = -0.030$, $SE = 0.062$, $z = -0.477$, $p = 0.634$). These results indicate that previously reading *Alice's Adventures in Wonderland* or watching a movie adaptation of the novel did not significantly relate to performance on this task.

2.3.2 Non-speech Stimuli—Two types of harmonic tone sequences (sampling rate = 22,500 Hz) were created. One tone sequence consisted of standard tones mixed with frequency deviant tones, and another consisted of standard tones mixed with duration deviant tones. The standard tones in both types of tone sequences were always comprised of three sinusoidal partials of 500, 1000, and 1500 Hz and were 75 ms in duration (including 5 ms rise and fall times). The intensity of the second and third partials was lower than that of

the first partial by 3 and 6 dB, respectively. The frequency deviant tones were 63% higher in frequency relative to the standard (partials: 800, 1600, 2400 Hz). The duration deviants were all 200 ms in duration. The probability of the deviant tones was the same for frequency and duration deviants and varied from ~13% to ~19% across trials. In a given sequence, the first 10 tones would always consist of the standard tones. The interval between consecutive tones was randomized between 300 to 500 ms.

2.3.3 Mixing speech and non-speech stimuli—

Both speech and non-speech stimuli were equated for root-mean-square (RMS) amplitude at 65 dB Sound Pressure Level (SPL). As shown in Figure 1A, each speech segment was mixed with a randomly selected tone sequence that either had a deviant that differed in duration or frequency. The duration of the tone sequences matched the duration of the corresponding narrative segments.

2.4 Procedures

Stimulus presentation was controlled by E-Prime 2.0.10 software (Schneider et al., 2002). Each participant listened to all 60 narrative segments (mixed with harmonic tone sequences) through two different conditions (30 unique narrative segments per condition): (1) an attend speech condition, and an (2) attend tone condition (Figure 1). Before the commencement of each condition, the task and EEG recording procedures were explained to the participant. In both conditions, participants were instructed to maintain visual fixation on a crosshair centered on the screen and to refrain from all other extraneous movements for the duration of each trial.

In the attend speech condition, participants were instructed to attend to the story narrative and ignore the tone sequence. In the attend tone condition, participants were instructed to attend to the tone sequence and ignore the story narrative. To ensure task compliance (i.e., effective modulation of attention to speech across conditions) and to measure comprehension, participants were required to answer four multiple-choice questions, two for the story narrative and two for the harmonic tone sequence after each trial. Participants had unlimited time to answer questions related to the story narrative (e.g., “What does Alice pick up while falling through the well?” and “How many miles does Alice think she has fallen?”). Likewise, participants had unlimited time to answer questions about the tone sequence (e.g., “How many tones were different?” and “In what way were the tones different?”). In the attend speech condition, the participants were instructed that the goal of the task was to get all questions correct for the questions related to the story and that it was not necessary to get questions related to the tone sequence correct. In the attend tone condition, the participants were instructed that the goal of the task was to get questions related to the tone sequence correct and that it was not necessary to get the questions related to the narrative correct. The order of conditions presented was counterbalanced across participants.

2.5 EEG Acquisition and Preprocessing

During EEG recordings, participants comfortably sat in a dark, acoustically shielded booth. The mixed speech and non-speech stimuli were binaurally presented via insert earphones (ER-3; Etymotic Research, Elk Grove Village, IL). EEG responses were recorded with 64 actiCAP active electrodes (Brain Products, Gilching, Munich, Germany) secured in an

elastic cap (EasyCap; www.easycap.de). Electrodes were placed on the scalp according to the International 10–20 system (Klem et al., 1999). A common ground was placed at the Fpz electrode site. EEG responses were amplified and digitized with a BrainVision actiCHAMP amplifier and recorded with PyCorder 1.0.7 software (Brain Products, Gilching, Munich, Germany) at a sampling rate of 25 kHz. Contact electrode impedance was less than 20 kΩ across all experimental conditions.

The EEG data were preprocessed off-line with BrainVision Analyzer 2.0 (Brain Products, Gilching, Germany). In line with similar EEG studies on neural tracking of the speech (Di Liberto et al., 2015), responses were bandpass filtered from 1 to 15 Hz, using a zero phase-shift Butterworth filter (12 dB/octave, zero phase shift), and referenced to the average of the two mastoid electrodes: TP9 and TP10. The EEG data were segmented into epochs that were time-locked to the onset of the mixed narrative and harmonic tone sequence stimuli. The duration of the epochs matched the respective auditory stimuli (~60 sec). EEG data were then downsampled to 128 Hz to reduce processing time. Independent component analysis (ICA) was performed on EEG responses in both conditions for each participant using the restricted Infomax algorithm. Ocular artifacts were identified and removed via visual inspection of their topographical distribution and activation pattern (time course). The remaining components were back-projected to the EEG electrode space. The EEG data was further downsampled to 64 Hz to improve computational efficiency for the neural tracking analysis.

2.6 Neural Tracking of the Speech Envelope

In line with previous investigations (Crosse et al., 2016; Di Liberto et al., 2015; Di Liberto & Lalor, 2017), we used a model-based analysis to quantify the extent to which the neural responses from native and non-native listeners reflected differences in neural tracking of the speech envelope. To this end, we conducted a regularized linear regression analysis to fit a TRF model that linearly mapped the continuous stimulus envelope, for a given trial, onto the neural response recorded at each EEG channel. TRF models imitate the impulse response of the neural system (i.e., how the system encodes information) and, in turn, are easily interpreted from a neural encoding/neurophysiological perspective (Haufe et al., 2014; Weichwald et al., 2015). We implemented TRF analyses using the multivariate temporal response function (mTRF) MATLAB (The MathWorks, Natick, MA) toolbox which is publicly available (https://sourceforge.net/projects/aespa/files/v1.4) (Crosse et al., 2016)

2.6.1 Extraction of the Amplitude Envelope—The multi-band speech envelope was extracted using Hilbert decomposition of the output of 16 gammatone filters spaced uniformly on an equivalent rectangular band scale in the frequency range of 250 through 8000 Hz. This was implemented using the Auditory Toolbox (Slaney, 1998). The amplitude of these multiband envelopes was raised to a power of 0.6 to mimic the compression by the inner ear (Decruy et al., 2019; Vanthornhout et al., 2018). These multiband envelopes were downsampled from 22,500 Hz to 64 Hz to match the sampling rate of the EEG.

2.6.2 Evaluation of the TRF and Estimating Neural Tracking Metrics—The EEG and multiband amplitude envelope of the speech stimuli across both attend speech and attend

tone conditions were used to obtain the temporal response functions via multivariate linear ridge regression (Figure 2). The temporal response functions were obtained over a restricted time lag window of −100 to 450 ms (Crosse et al., 2016; Di Liberto et al., 2015; Di Liberto & Lalor, 2017). The TRFs were obtained using a 30-fold cross-validation approach, where the TRF was trained on 29 of the 30 trials for each condition separately. Then, TRFs averaged over the 29 trials were used to predict the EEG data from the remaining trial (unseen data). This process recurred until the EEG responses were predicted from all trials. To prevent overfitting of the TRFs, we conducted a parameter search (over the range $2^{1,2,3,\ldots 10}$) to identify the ridge regression parameter that optimized EEG prediction for in each group (Crosse et al., 2016). The optimal ridge parameter was chosen based on the best model fit in each condition, collapsed across the groups. EEG prediction accuracies were calculated at each electrode via a correlation coefficient (Pearson's r) between the actual and predicted EEG data. The correlation coefficients were then averaged across the 30 trials, and a single neural tracking metric was derived by averaging these correlations across the 62 electrodes. These neural tracking metrics are essentially the similarity between the speech stimulus envelope and the EEG recordings. Greater values indicate better neural tracking of the speech stimulus envelopes. The TRFs complemented the neural tracking metrics and informed about the temporal dynamics of neural tracking. The neural tracking metrics and temporal response functions were compared between the two groups and conditions to assess the effect of language experience and attention. The chance level for the neural tracking metrics was estimated by predicting EEG to a mismatched stimulus based on the best TRF model obtained above. This process was reiterated for 1000 random permutations (with replacement) of the stimulus and EEG data.

2.6.3 Cortical Auditory Evoked Potentials (CAEPs) to the Tone Stimuli—

Unlike the neural tracking to the speech envelope, cortical responses to the tone sequences were extracted in the conventional manner by time-locked averaging of the EEG to the onset of individual tones in the sequence. This was different than speech, as speech was continuously presented and was dynamically fluctuating overtime, which warranted a TRF based tracking analytic approach rather than conventional event-related potential approach. However, the tones in a sequence were presented repetitively and could be analyzed in the conventional manner as other event-related potentials. Thus, we leveraged the most appropriate procedures available to analyze the responses to the speech tracks and tone sequences.

The EEG epochs time-locked to all the standard stimuli were extracted using a time window from −100 to 450 ms with 0 ms corresponding to the tone onset. The CAEP epochs were baseline corrected and averaged across trials for the two conditions separately. We also evaluated the TRFs to the tone stimuli to supplement the CAEP analyses. The TRFs showed morphology and topographies similar to the CAEPs (see Supplementary Figure 2). Indeed, TRFs are close approximations to the actual CAEPs, but one downside of this analytic approach is that the results are affected by smoothing -- which is an integral component of the estimation process (Crosse et al. 2016). In turn, CAEPs seem to be best suited for analyzing cortical responses to repetitive presentation of a given stimuli (which is how the

tone stimuli were presented in the current investigation). Therefore, the tone TRF analyses were conducted as a supplement to the main CAEP analyses and results.

2.7 Statistical Analyses

Linear mixed-effects models were implemented using the *lme4* package in R version 3.2.5. P-values for model parameters were obtained with the *lmerTest* package using the Satterthwaite method for estimating degrees of freedom. Fixed effects included language group (non-native or native), condition (attend speech and attend tone), and their interactions. To account for baseline differences across subjects, we included a by-subject intercept as a random effect. Unless otherwise specified, in all models, the reference levels were the native language group and attend speech condition. Similar analyses using linear mixed-effects regression in MATLAB was performed on the time-course (every time-point in the waveform) of the temporal response function and the CAEPs to assess the effects of group and attention. The multiple comparison problem for the high dimensional data was controlled by using cluster-based permutations (Maris & Oostenveld, 2007). The main effects of group are illustrated by averaging the waveforms in the two attention conditions within the listener groups. Similarly, the main effects of attention are illustrated by averaging the waveforms of the two listener groups within each attention condition. The latency regions with statistically significant main effects after controlling for multiple comparisons are represented as gray shaded regions on the waveforms. The main effect of group and attention across the different electrodes are illustrated as average differences in topography for each comparison across the different latency regions. The sum of the t-values (t_{clust}) within each significant cluster is also reported.

3. RESULTS

3.1 Behavioral Results

3.1.1 Speech—Both native and non-native listeners demonstrated higher mean speech comprehension accuracy and slower reaction times in the attend speech condition relative to attend tone (ignore speech) condition. The native participants exhibited significantly higher speech comprehension accuracy and faster reaction times relative to the non-native participants (Figure 3). Confirming this observation, a linear mixed-effects model revealed a significant main effect of listener group [accuracy: $\beta = -22.50$, $SE = 3.47$, χ^2 (1) = 112.05, $p < 0.001$; RT: $\beta = 3577.45$, $SE = 693.35$, χ^2 (1) = 13.31, $p = 0.0003$], indicating that non-native participants exhibited significantly lower accuracy and slower reaction times answering the speech multiple choice questions relative to the native participants. A significant main effect of condition was also found [accuracy: $\beta = -18.13$, $SE = 2.95$, χ^2 (1) = 112.05, $p < 0.001$; RT: $\beta = 661.86$, $SE = 503.44$, χ^2 (1) = 4.37, $p = 0.036$], indicating that in general, participants exhibited lower speech comprehension accuracy and faster reaction times in the attend tone condition relative to the attend speech condition. There was also a significant interaction effect between language group and condition [accuracy: $\beta = -13.75$, $SE = 4.17$, χ^2 (1) = 10.85, $p = 0.001$; RT: $\beta = -2812.86$, $SE = 711.97$, χ^2 (1) = 15.61, $p < 0.001$]. Post hoc contrasts extracted from the linear mixed-effects model indicated that native language participants had significantly higher speech comprehension accuracy and faster reaction times compared to non-native participants in the attend speech condition [accuracy:

$\beta = 22.50$, $SE = 3.47$, $t = 6.48$, $p < 0.001$; reaction time: $\beta = -3577$, $SE = 693$, $t = -5.16$, $p < 0.001$]. The difference in behavioral accuracy between native and non-native participants was most pronounced in the attend tone condition [$\beta = 36.20$, $SE = 3.47$, $t = 10.44$, $p < 0.001$]; with no difference found in speech reaction times to speech in the attend tone condition [$\beta = -765$, $SE = 693$, $t = -1.10$, $p = 0.276$].

3.1.2 Non-speech Stimuli—Both native and non-native participants additionally demonstrated higher mean behavioral accuracy on the questions about the tone sequence in the attend tone (ignore speech) condition relative to attend speech condition (Figure 3). For the tone multiple-choice questions, we found a significant main effect of condition [accuracy: $\beta = 19.79$, $SE = 3.57$, χ^2 (1) = 86.47, $p < 0.001$; RT: $\beta = 371.03$, $SE = 231.52$, χ^2 (1) = 13.44, $p < 0.001$], with higher non-speech behavioral accuracy and slower reaction times in the attend tone condition (when participants were instructed to attend to the tone sequence), relative to the attend speech condition (when participants were instructed to ignore the tone sequence). The main effect of group [accuracy: $\beta = -8.96$, $SE = 4.07$, χ^2 (1) = 2.71, $p = 0.099$; RT: $\beta = -131.58$, $SE = 269.54$, χ^2 (1) = 0.21, $p = 0.648$] and the interaction effect between group and condition [accuracy: $\beta = 7.39$, $SE = 5.05$, χ^2 (1) = 2.14, $p = 0.143$; RT: $\beta = 458.44$, $SE = 327.41$, χ^2 (1) = 1.96, $p = 0.161$] were not statistically significant.

3.2 Effect of Attention and Language Experience on Neural Tracking of the Speech Envelope

Although non-native listeners exhibited lower behavioral accuracy on the speech questions both in the attend speech and attend tone (ignore speech) conditions, they exhibited greater neural tracking of the speech envelope in both conditions (Figure 4). Specifically, there was a significant main effect of language group [$\beta = 0.012$, $SE = 0.005$, χ^2 (1) = 5.53, $p = 0.019$], indicating that the non-native participants exhibited greater neural tracking of the speech envelope relative to native participants. The main effect of condition was also significant [$\beta = -0.005$, $SE = 0.003$, χ^2 (1) = 10.64, $p = 0.001$], indicating that neural tracking of the speech envelope was higher in the attend speech condition relative to the attend tone condition. The interaction between condition and group was not statistically significant [$\beta = -0.005$, $SE = 0.004$, χ^2 (1) = 1.44, $p = 0.230$], indicating that the pattern of enhanced neural tracking of the speech envelope observed in non-native listeners relative to native listeners did not differ across attend speech and attend tone conditions.

Complementary to this, the effects of group and condition on the temporal course of neural tracking was assessed by statistical comparison of the TRFs. The TRFs are the time series of regression (beta) weights that explain the extent to which the stimulus envelope is mapped onto the EEG at different time lags. Higher absolute beta weights indicate better neural tracking. The TRFs showed three prominent peaks consistent with previous studies (Crosse et al., 2016). The results of the statistical comparison of TRFs showed significant main effects (significant time regions and electrodes marked based on cluster corrected significance level) of both language group and attention condition (Figure 5A). In Figure 5A, the TRFs are pooled across attention condition to show the main effect of group and pooled across groups to show the main effect of condition. Comparison across conditions

showed that TRFs in the attend speech condition showed significantly larger beta weights than in the attend tone condition. These differences were seen in three distinct latency regions consistent with the three TRF peaks that also differed in their topographies (78 to 93 ms, t_{clust} = 97.5; 140 to 156 ms, t_{clust} = 73.9, $p < 0.01$; 250 to 256 ms, t_{clust} =166.7, $p < 0.01$). The group effects showed different patterns of scalp-topography. Comparing TRFs across listener groups, the TRFs in the non-native listener group showed higher beta weights than the native listener group. This is consistent with higher neural tracking (r-values) in the non-native listener group. These differences were seen at two distinct latency regions, which also differed in topography across the scalp (109 to 156 ms, t_{clust} = 251.4, $p < 0.01$; 218 to 234 ms, t_{clust} =39.16, $p < 0.01$). Additionally, the interaction effect between group and condition was not statistically significant, suggesting that the pattern of group differences in the TRFs did not vary across conditions. This suggests that the difference in cortical tracking obtained in the two listener groups is not significant because of differences in attentional effects on neural tracking, but rather are driven by overall group differences. The absence of group effects on the attentional component is also shown in figure 5A as the difference waveforms of attend speech - attend tone within each group.

3.3 Relationship Between Neural tracking of Speech and Behavior

Next, we examined the extent to which neural tracking of the speech envelope related to speech comprehension accuracy and reaction times that were measured while participants answered story comprehension questions. Spearman's rank correlation coefficients indicated that story comprehension accuracy was not significantly associated with neural tracking in the attend speech condition but was significantly negatively correlated with neural tracking in the attend tone condition across all participants [attend speech: $r = -0.333$, $p = 0.077$; attend tone: $r = -0.385$, $p = 0.039$]. Speech comprehension accuracy was not significantly correlated with neural speech tracking within the native listener group [attend speech: $r = 0.093$, $p = 0.740$; attend tone: $r = -0.378$, $p = 0.164$], or within the non-native listener group [attend speech: $r = -0.002$, $p = 0.994$; attend tone: $r = -0.210$, $p = 0.471$]. Speech reaction times pooled across all participants significantly correlated with neural tracking in the attend speech condition [$r = 0.403$, $p = 0.030$], but not in the attend tone condition [$r = 0.111$, $p = 0.564$]. Speech reaction times were not significantly correlated with neural tracking within the native listener group [attend speech: $r = 0.060$, $p = 0.832$; attend tone: $r = -0.239$, $p = 0.389$], or within the non-native listener group [attend speech: $r = 0.134$, $p = 0.648$; attend tone: $r = 0.116$, $p = 0.693$]. These findings suggest that individual differences in neural tracking do not relate to behavioral performance.

3.4 Effect of Attention and Language Experience on the Cortical Auditory Evoked Potential to Tone Stimuli

The effects of group and condition on the CAEP to the tone was statistically assessed (Figure 5B). The main effect of attention was seen at several electrodes in the latency region of the 70–140 ms (t_{clust} = 685, $p < 0.01$). The attention effects were seen as higher negativity in the attend tone condition (when listeners were instructed to attend to the tone sequence) than the attended speech condition (when listeners were instructed to ignore the tone sequence). There was also a significant main effect of listener group on the CAEPs in the latency region 86–93 ms (t_{clust} = 81.22, $p < 0.01$). The group effect was observed as higher

negativity in the non-native listener group compared to the native listener group. Additionally, the interaction effect between group and condition was not statistically significant, suggesting that the pattern of group differences in the CAEPs did not differ across attentional conditions. This suggested that the difference in CAEPs obtained in the two listener groups is not simply due to differences in attentional effects related to the sensory encoding of the harmonic tone complex. The absence of group effects on the attentional component is also shown in figure 5B as the difference waveforms of attend speech - attend tone.

4. DISCUSSION

We examined neural tracking of the speech envelope while native and non-native participants listened to an English narrative simultaneously presented with a harmonic tone sequence (non-speech stimuli). This experimental design created a challenging listening environment without the confound of a linguistic distractor. Both listener groups showed greater speech comprehension in the attend speech relative to the attend tone condition. Compared to native listeners, non-native listeners exhibited lower speech comprehension accuracy in both the attend speech and attend tone (ignore speech) conditions. Yet, non-native listeners showed *enhanced* neural tracking of the speech envelope in both attend speech and attend tone conditions, relative to native listeners. Additionally, non-native listeners exhibited enhanced encoding of the non-speech tone stimuli compared to native listeners in both conditions. Taken together, these results suggest that, compared to native listeners, non-native listeners employ a different listening strategy when attending to speech and non-speech stimuli in challenging listening environments that is manifested as a non-specific neural amplification of low-level acoustic cues.

Irrespective of the listener group, neural tracking of the speech envelope was greater when participants were attending to the speech signal. These results are consistent with a large body of evidence demonstrating that attention increases the neural tracking of the speech envelope (Ding & Simon, 2012; Golumbic et al., 2013; Rimmele et al., 2015; Vanthornhout et al., 2019). In contrast to the prior studies, we were able to examine attention effects without the confound of linguistic interference from the masker. The findings of the current investigation indicate that attention-induced enhancement in the neural tracking of the speech envelope is a reliable marker, present even where there is minimal informational overlap between the target speech signal and the masker.

A key finding from the current study is that non-native listeners exhibited heightened neural tracking of the speech envelope relative to native listeners across both attentional conditions. That is, non-native listeners exhibited higher neural tracking even when they were instructed to ignore the speech stimuli and attend to a non-speech signal. These results are consistent with prior studies that have demonstrated enhanced neural tracking of the speech envelope in non-native listeners (Song & Iverson, 2018; Zou et al., 2019). For example, naïve listeners for whom speech stimuli are unintelligible show greater neural tracking of the speech envelope relative to native listeners (Zou et al., 2019). In Zou et al. (2019), naïve listeners performed poorer than native listeners when detecting repeated sound segments in a continuous speech signal, even though they exhibited enhanced neural tracking of the speech

envelope relative to native listeners. This pattern is strikingly similar to the results from the current study.

The CAEPs to the tone sequence (Näätänen & Picton, 1987) also showed effects of attention in both groups. CAEPs to the tone sequence showed stronger negativity in 78–132 ms latency regions in attend tone condition compared to the attend speech condition. Prior studies have referred to the greater negativity to the selectively attended stimulus as the Nd wave or the slow negative shift. The Nd wave has been previously linked to selective attention and orientation (Michie et al., 1990; Näätänen & Michie, 1979; Parasuraman, 1980). The results of the current investigation suggest that selective attention effects are also evidenced in the CAEPs (attend tone > attend speech). This attention effect in the CAEP to the non-speech stimuli was similar in both listener groups. Interestingly, the CAEPs in non-native listeners showed greater negativity than native listeners in the latency region of 86–93 ms. Consistent with previous literature, non-native listeners exhibited enhanced attention to non-speech stimuli irrespective of the attentional task (Strait et al., 2014). This may be because non-native speech *may not* be as distracting as native speech, for non-native listeners (Mahajan & McArthur, 2011). To further support this interpretation of the current results, native listeners, compared to non-native listeners, seemed to be more distracted by the speech of their native language in the attend tone condition (which is evidenced by greater speech comprehension accuracy in the attend tone condition). A recent study supporting this interpretation, showed that familiar and more intelligible competing signals are more distracting than unfamiliar or less intelligible competing signals (Olguin et al., 2018). In the current investigation, the English language was less familiar to the non-native listeners and therefore the speech signal could have been less distracting. In turn, non-native listeners might have been able to allocate greater attentional resources to the tone sequence (non-speech stimuli) in both attentional conditions. The stronger negativity in the CAEPs and higher neural tracking on the speech envelope in the non-native listeners thus suggests that non-native listeners show heightened attention to low-level acoustic cues in both the speech and non-speech stimuli. The presence of effects of attention and language experience exhibited by both the speech TRFs and the CAEPs should not be interpreted as emerging from a common source or neural generator site, considering the differences in response characteristics i.e., latency, morphology as well as differential sensitivity to acoustic cues in speech and non-speech stimuli (Aiken & Picton, 2008; Easwar et al., 2012; Swink & Stuart, 2012). Such an interpretation and direct comparison of sources underlying CAEPs to tones and speech TRFs is challenging given the limited spatial resolution offered by EEG.

We also performed an analysis using TRF-based approach for the tone onsets as a supplement to the CAEP analysis (see Supplementary Results and Supplementary Figure 2). The TRF-based neural correlation coefficients (i.e., Pearson's r correlation coefficients calculated between the actual and predicted EEG data based on tone onsets) did not show effects of attention or language experience. In comparing the temporal course of the tone TRFs to that of the CAEPs, both analyses yielded attention effects (attend tone > attend speech); however, the onset of the statistical differences occurred at an earlier latency in the CAEPs than the tone TRFs. Although the CAEP results showed group effects (non-native > native), the tone TRFs did not reveal significant group differences. This discrepancy might be a result of the TRF analysis technique, which in essence, is an approximation of the

actual CAEPs, and involves smoothing which is integral to the TRF procedure. This in turn could result in a potential loss of information. Because the tone stimuli were presented repetitively and could be analyzed in the conventional manner as other event-related potentials via CAEP, we leveraged this approach over the TRF approach to examine neural responses to the tone stimuli in native and non-native listeners.

What is the functional role of enhanced neural tracking of the speech envelope? Prior work has suggested several explanations (Ding & Simon, 2014). One set of explanations consistent with domain-general bottom-up processing models indicates that the speech envelope is a crucial cue for speech segmentation of continuous speech. Enhanced tracking of the envelope may aid in speech segmentation and, consequently, phonemic parsing. Our results (non-native > native) are clearly not consistent with this explanation. The second set of explanations are consistent with interactive processing models. That is, neural tracking of the envelope is influenced by group differences (native vs. non-native) in the efficiency of top-down predictive processes. Our results do not completely support the interactive processing models either. Indeed, increased reliance on the envelope could be necessitated by limited language proficiency in non-native listeners. However, non-native listeners also increase the tracking of the speech envelope when specifically instructed to ignore the speech stream. We posit a more integrative account: the incomplete language model (Mattys et al., 2010, 2012) in non-native listeners, which necessitates a listening mode that primes greater encoding of bottom-up acoustic features. However, the heightened processing of bottom-up features may not be goal-directed or efficient. Additionally, non-native listeners may also need to suppress the non-target native language, more so than native, monolingual listeners (Hilchey & Klein, 2011). Native listeners might expend fewer resources on these features and instead efficiently utilize higher-level linguistic cues (e.g., lexical plausibility) to process continuous speech efficiently. It is worthy to note that the results from our experiment cannot be directly compared to existing literature (Ding & Simon, 2011; Golumbic et al., 2013; J. R. Kerlin et al., 2010) on neural tracking of speech during selective attention. In the previous literature selective attention has been evaluated when a listener paid attention to a one speech stream while ignoring another speech stream. Such a paradigm usually leads to a listener being able to understand only one speech stream and very poor understanding of the ignored speech stream and involves all sorts of energetic and informational (linguistically loaded) masking effects. In our study however, there was only one speech stream and one non-speech stream that was a tone sequence. The energetic masking effects of the tone sequence on speech would be minimal because of the presence of only three frequency components in each complex -tone in the tone sequence. The informational masking effects are also minimal. Thus, it is possible that the participants (especially the native listeners) were able to perceive enough information from the story narrative even while they were selectively attending to the tone sequence. Such use of speech and non-speech stimuli in future studies could be leveraged to understand the influence of selective attention on neural tracking of speech without the confounds of energetic or informational masking.

4.1 Conclusions

The ability to selectively attend to an incoming speech signal amid competing ambient sounds is a significant sensory and cognitive challenge, especially for listeners who must attend to and understand the speech of their non-native language (Lecumberri et al., 2010; Scharenborg & van Os, 2019). Consistent with the incomplete language model (Mattys et al., 2010, 2012), the current investigation indicates that during speech comprehension under a cognitively demanding task, non-native listeners with limited language proficiency for the attended speech signal tend to rely more on acoustic cues. This is reflected by enhanced neural tracking of the speech envelope. Our findings are in support of an emerging view that although attention to a target speech signal enhances neural tracking of the speech envelope, this mechanism itself may not confer speech comprehension advantages. In non-native listeners, enhanced neural tracking of the speech envelope may indicate a processing mode that non-specifically enhances bottom-up features in the incoming stimulus stream.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Preliminary findings from this original investigation were presented as a poster presentation at the 2017 International Conference on Auditory Cortex (ICAC), where helpful feedback from peers was received. The authors would like to thank Zilong Xie for the generation of the harmonic tone sequence and assistance with the initial application of the mTRF toolbox. We also thank the members of the SoundBrain Laboratory for assistance with participant recruitment, data collection, and data preprocessing. This work was also supported in part by the resources provided through the University of Pittsburgh Center for Research Computing (https://doi.org/10.1002/open.201700046).

FUNDING

This work was supported by the National Institute On Deafness and Other Communication Disorders of the National Institutes of Health [R01DC015504, R01DC013315 (BC)].

REFERENCES

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, & Merzenich MM (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proceedings of the National Academy of Sciences, 98(23), 13367–13372.

Aiken SJ, & Picton TW (2008). Human cortical responses to the speech envelope. Ear and Hearing, 29(2), 139–157. [PubMed: 18595182]

Astheimer LB, Berkes M, & Bialystok E (2016). Differential allocation of attention during speech perception in monolingual and bilingual listeners. Language, Cognition and Neuroscience, 31(2), 196–205.

Bidelman GM, & Dexter L (2015). Bilinguals at the “cocktail party”: Dissociable neural activity in auditory–linguistic brain regions reveals neurobiological basis for nonnative listeners’ speech-in-noise recognition deficits. Brain and Language, 143, 32–41. [PubMed: 25747886]

Bidelman GM, Gandour JT, & Krishnan A (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. Journal of Cognitive Neuroscience, 23(2), 425–434. [PubMed: 19925180]

Bregman AS (1994). Auditory scene analysis: The perceptual organization of sound. MIT press.

Brouwer S, Van Engen KJ, Calandruccio L, & Bradlow AR (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. The Journal of the Acoustical Society of America, 131(2), 1449–1464. [PubMed: 22352516]

Cooke M, Garcia Lecumberri M, & Barker J (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. The Journal of the Acoustical Society of America, 123(1), 414–427. [PubMed: 18177170]

Crosse MJ, Di Liberto GM, Bednar A, & Lalor EC (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. Frontiers in Human Neuroscience, 10, 604. [PubMed: 27965557]

Cutler A, Weber A, Smits R, & Cooper N (2004). Patterns of English phoneme confusions by native and non-native listeners. The Journal of the Acoustical Society of America, 116(6), 3668–3678. [PubMed: 15658717]

Decruy L, Vanthornhout J, & Francart T (2019). Tracking the speech envelope of target versus competing speech: Normal-hearing versus hearing impaired listeners. Auditory EEG Signal Processing (AESoP) symposium, Date: 2019/09/16–2019/09/18, Location: Leuven, Belgium.

Di Liberto GM, & Lalor EC (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. Hearing Research, 348, 70–77. [PubMed: 28246030]

Di Liberto GM, O'Sullivan JA, & Lalor EC (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology, 25(19), 2457–2465. [PubMed: 26412129]

Ding N, Patel AD, Chen L, Butler H, Luo C, & Poeppel D (2017). Temporal modulations in speech and music. Neuroscience & Biobehavioral Reviews, 81, 181–187. [PubMed: 28212857]

Ding N, & Simon JZ (2011). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophysiology, 107(1), 78–89. [PubMed: 21975452]

Ding N, & Simon JZ (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences, 109(29), 11854–11859.

Ding N, & Simon JZ (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. Frontiers in Human Neuroscience, 8, 311. [PubMed: 24904354]

Doelling KB, Arnal LH, Ghitza O, & Poeppel D (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage, 85, 761–768. [PubMed: 23791839]

Easwar V, Glista D, Purcell DW, & Scollie SD (2012). The effect of stimulus choice on cortical auditory evoked potentials (CAEP): Consideration of speech segment positioning within naturally produced speech. International Journal of Audiology, 51(12), 926–931. [PubMed: 22916693]

Flege JE, & Hillenbrand J (1986). Differential use of temporal cues to the/s/–/z/contrast by native and non-native speakers of English. The Journal of the Acoustical Society of America, 79(2), 508–517. [PubMed: 3950204]

Fu Q-J, Zeng F-G, Shannon RV, & Soli SD (1998). Importance of tonal envelope cues in Chinese speech recognition. The Journal of the Acoustical Society of America, 104(1), 505–510. [PubMed: 9670541]

Fuglsang SA, Dau T, & Hjortkjær J (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. NeuroImage, 156, 435–444. [PubMed: 28412441]

Giraud A-L, & Poeppel D (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience, 15(4), 511. [PubMed: 22426255]

Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, & Simon JZ (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." Neuron, 77(5), 980–991. [PubMed: 23473326]

Hambrook DA, & Tata MS (2014). Theta-band phase tracking in the two-talker problem. Brain and Language, 135, 52–56. [PubMed: 24911919]

Hammill DD, Brown VL, Larsen SC, & Wiederholt JL (2007). Test of adolescent and adult language. Pro-Ed Austin, TX.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, & Bießmann F (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87, 96–110. [PubMed: 24239590]

Hilchey MD, & Klein RM (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. Psychonomic Bulletin & Review, 18(4), 625–658. [PubMed: 21674283]

Joris P, Schreiner C, & Rees A (2004). Neural processing of amplitude-modulated sounds. Physiological Reviews, 84(2), 541–577. [PubMed: 15044682]

Kaufman AS, & Kaufman NL (2004). Kaufman Brief Intelligence Test—Second Edition (KBIT-2). Pearson Publishers https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Cognition-%26-Neuro/Non-Verbal-Ability/Kaufman-Brief-Intelligence-Test-%7C-Second-Edition/p/100000390.html

Kerlin JR, Shahin AJ, & Miller LM (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party." Journal of Neuroscience, 30(2), 620–628. 10.1523/JNEUROSCI.3631-09.2010 [PubMed: 20071526]

Kerlin Jess R, Shahin AJ, & Miller LM (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party." Journal of Neuroscience, 30(2), 620–628. [PubMed: 20071526]

Klem GH, Lüders HO, Jasper H, & Elger C (1999). The ten-twenty electrode system of the International Federation. Electroencephalogr Clin Neurophysiol, 52(3), 3–6.

Lecumberri MLG, Cooke M, & Cutler A (2010). Non-native speech perception in adverse conditions: A review. Speech Communication, 52(11–12), 864–886.

Li P, Zhang F, Tsai E, & Puls B (2014). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. Bilingualism: Language and Cognition, 17(3), 673–680.

Mahajan Y, & McArthur G (2011). The effect of a movie soundtrack on auditory event-related potentials in children, adolescents, and adults. Clinical Neurophysiology, 122(5), 934–941. [PubMed: 20869913]

Maris E, & Oostenveld R (2007). Nonparametric statistical testing of EEG-and MEG-data. Journal of Neuroscience Methods, 164(1), 177–190. [PubMed: 17517438]

Mattys SL, Carroll LM, Li CK, & Chan SL (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. Speech Communication, 52(11–12), 887–899.

Mattys SL, Davis MH, Bradlow AR, & Scott SK (2012). Speech recognition in adverse conditions: A review. Language and Cognitive Processes, 27(7–8), 953–978.

Mayo LH, Florentine M, & Buus S (1997). Age of second-language acquisition and perception of speech in noise. Journal of Speech, Language, and Hearing Research, 40(3), 686–693.

McClelland JL, & Elman JL (1986). The TRACE model of speech perception. Cognitive Psychology, 18(1), 1–86. [PubMed: 3753912]

Michie PT, Bearpark HM, Crawford JM, & Glue LC (1990). The nature of selective attention effects on auditory event-related potentials. Biological Psychology, 30(3), 219–250. [PubMed: 2282370]

Näätänen R, & Michie PT (1979). Early selective-attention effects on the evoked potential: A critical review and reinterpretation. Biological Psychology, 8(2), 81–136. [PubMed: 465623]

Näätänen R, & Picton T (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. Psychophysiology, 24(4), 375–425. [PubMed: 3615753]

Newman AJ, Tremblay A, Nichols ES, Neville HJ, & Ullman MT (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. Journal of Cognitive Neuroscience, 24(5), 1205–1223. [PubMed: 21981676]

Olguin A, Bekinschtein TA, & Bozic M (2018). Neural encoding of attended continuous speech under different types of interference. Journal of Cognitive Neuroscience, 30(11), 1606–1619. [PubMed: 30004849]

O'sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, & Lalor EC (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex, 25(7), 1697–1706. [PubMed: 24429136]

Pakulak E, & Neville HJ (2010). Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. Journal of Cognitive Neuroscience, 22(12), 2728–2744. [PubMed: 19925188]

Parasuraman R (1980). Effects of information processing demands on slow negative shift latencies and N100 amplitude in selective and divided attention. Biological Psychology, 11(3–4), 217–233. [PubMed: 7272394]

Rimmele JM, Golumbic EZ, Schröger E, & Poeppel D (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. Cortex, 68, 144–154. [PubMed: 25650107]

Rogers CL, Lister JJ, Febo DM, Besing JM, & Abrams HB (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. Applied Psycholinguistics, 27(3), 465–485.

Rosen S (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 336(1278), 367–373. [PubMed: 1354376]

Scharenborg O, & van Os M (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. Speech Communication.

Schneider W, Eschman A, & Zuccolotto A (2002). E-Prime 2.0 software. Psychology Software Tools Inc: Pittsburgh, PA, USA.

Shamma SA, Elhilali M, & Micheyl C (2011). Temporal coherence and attention in auditory scene analysis. Trends in Neurosciences, 34(3), 114–123. [PubMed: 21196054]

Shannon RV, Zeng F-G, Kamath V, Wygonski J, & Ekelid M (1995). Speech recognition with primarily temporal cues. Science, 270(5234), 303–304. [PubMed: 7569981]

Shinn-Cunningham BG (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186. [PubMed: 18396091]

Slaney M (1998). Auditory toolbox. Interval Research Corporation, Tech. Rep, 10(1998).

Song J, & Iverson P (2018). Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. Cognition, 179, 163–170. [PubMed: 29957515]

Steinschneider M, Nourski KV, & Fishman YI (2013). Representation of speech in human auditory cortex: Is it special? Hearing Research, 305, 57–73. [PubMed: 23792076]

Strait DL, Slater J, Abecassis V, & Kraus N (2014). Cortical response variability as a developmental index of selective auditory attention. Developmental Science, 17(2), 175–186. [PubMed: 24267508]

Swink S, & Stuart A (2012). Auditory long latency responses to tonal and speech stimuli. Journal of Speech, Language, and Hearing Research: JSLHR, 55(2), 447–459. 10.1044/1092-4388(2011/10-0364)

Vanthornhout J, Decruy L, & Francart T (2019). Effect of task and attention on neural tracking of speech. Frontiers in Neuroscience, 13, 977. [PubMed: 31607841]

Vanthornhout J, Decruy L, Wouters J, Simon JZ, & Francart T (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. Journal of the Association for Research in Otolaryngology, 19(2), 181–191. [PubMed: 29464412]

Weber-Fox C, Davis LJ, & Cuadrado E (2003). Event-related brain potential markers of high-language proficiency in adults. Brain and Language, 85(2), 231–244. [PubMed: 12735941]

Weichwald S, Meyer T, Özdenizci O, Schölkopf B, Ball T, & Grosse-Wentrup M (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage, 110, 48–59. [PubMed: 25623501]

Wong PC, Skoe E, Russo NM, Dees T, & Kraus N (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nature Neuroscience, 10(4), 420–422. [PubMed: 17351633]

Xie Z, Reetzke R, & Chandrasekaran B (2019). Machine Learning Approaches to Analyze Speech-Evoked Neurophysiological Responses. Journal of Speech, Language, and Hearing Research, 62(3), 587–601.

Zou J, Feng J, Xu T, Jin P, Luo C, Zhang J, Pan X, Chen F, Zheng J, & Ding N (2019). Auditory and language contributions to neural encoding of speech features in noisy environments. NeuroImage, 192, 66–75. [PubMed: 30822469]

HIGHLIGHTS

- Attention and language effects on neural speech and non-speech processing were studied.
- Attention modulated neural speech and non-speech processing for all listeners.
- Native listeners had relatively better speech comprehension.
- Non-native listeners had enhanced neural responses to speech and non-speech stimuli.
- Neural attentional enhancement in not speech-specific in non-native listeners.

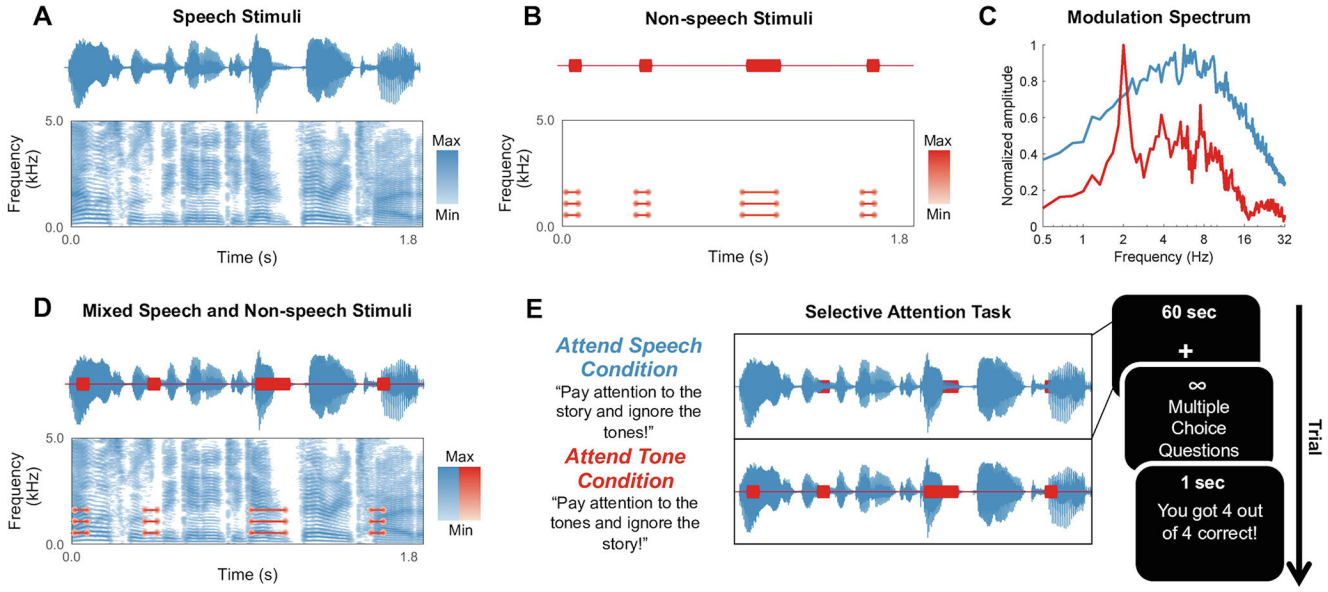

Figure 1. Experimental Methods.

Waveforms and spectrograms of a segment of (A) speech and (B) non-speech stimuli used in the study. (C) Modulation spectrum of a single track of the speech and non-speech stimuli. (D) Mixed speech and non-speech stimuli waveforms and spectrograms. (E) Trial procedure for attend speech and attend tone conditions. Each 60 s story segment was played while participants maintained visual fixation on a crosshair centered on the screen for the duration of each trial. In the attend speech condition, participants were instructed to attend to the story and ignore the tone sequence. Each condition consisted of 30 unique story segments mixed with a unique tone sequence. Participants were given unlimited time to respond to multiple choice questions about both the story and the tone sequence after listening to the speech-tone stimuli. Corrective feedback was presented for 1 s following participants' response.

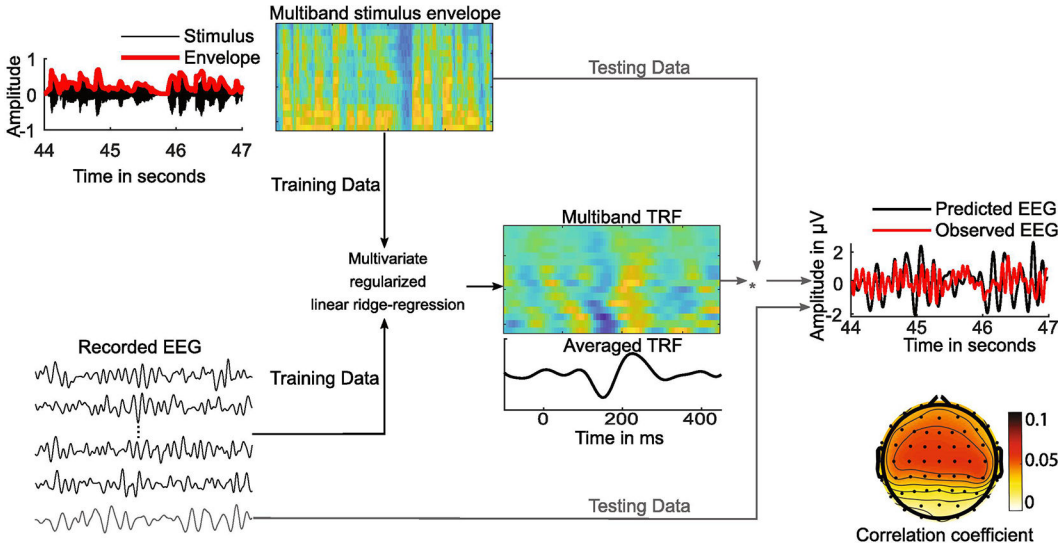

Figure 2. Temporal Response Function.

Used to estimate neural tracking of the speech envelope. This illustrates the multivariate regularized linear ridge regression using k-fold cross-validation to evaluate the neural tracking of speech envelope using the forward encoding model.

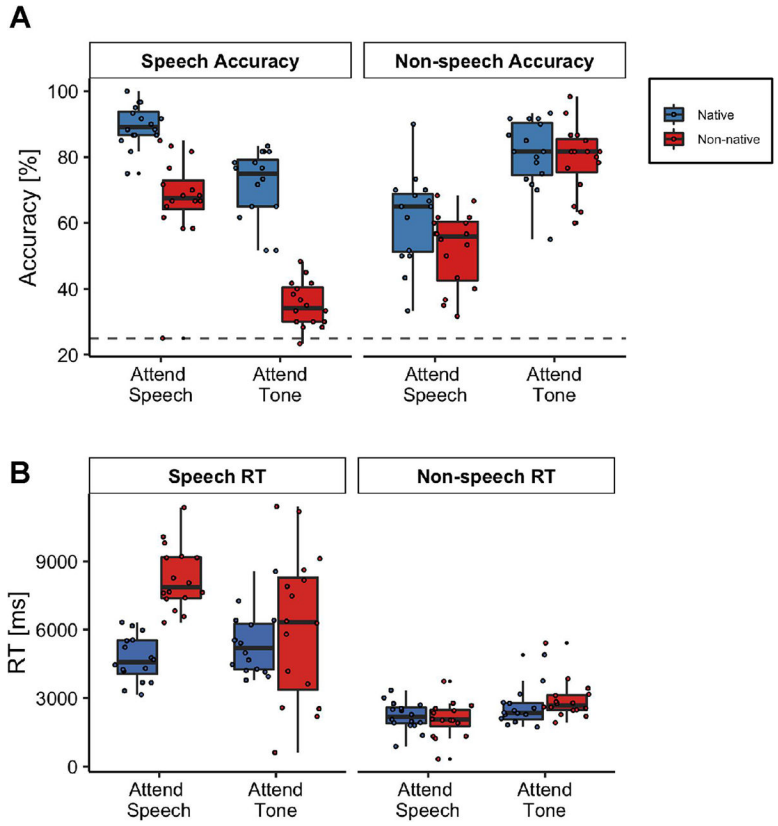

Figure 3. Behavioral Results.

Speech and non-speech behavior (A) accuracy and (B) reaction times. The center line on each box plot denotes the median accuracy or reaction time, the edges of the box indicate the 25th and 75th percentiles, and the whiskers extend to data points that lie within 1.5 times the interquartile range. Points outside this range represent outliers. The dashed line in (A) denotes chance level.

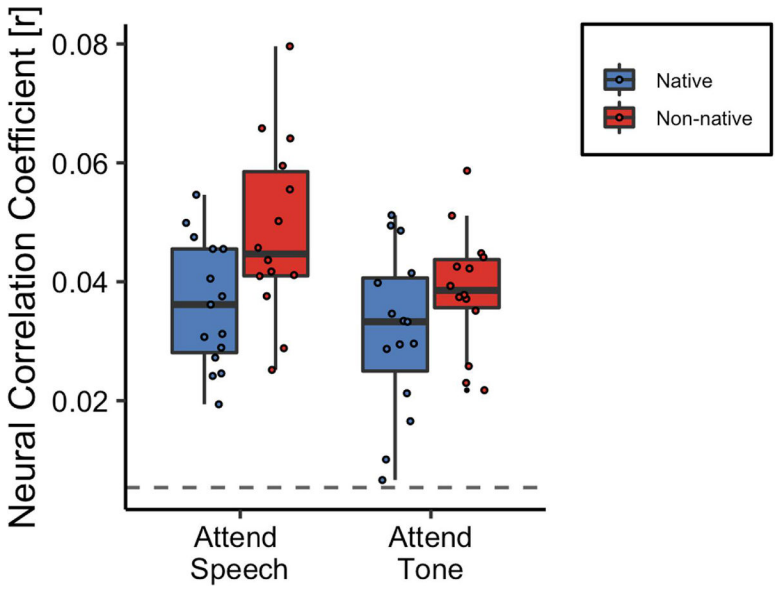

Figure 4. Neural tracking of the speech envelope is modulated by attention and language experience.

The center line on each box plot denotes the median neural correlation coefficient, the edges of the box indicate the 25th and 75th percentiles, and the whiskers extend to data points that lie within 1.5 times the interquartile range. Points outside this range represent outliers. The dashed line denotes chance estimates for the neural correlation coefficient.

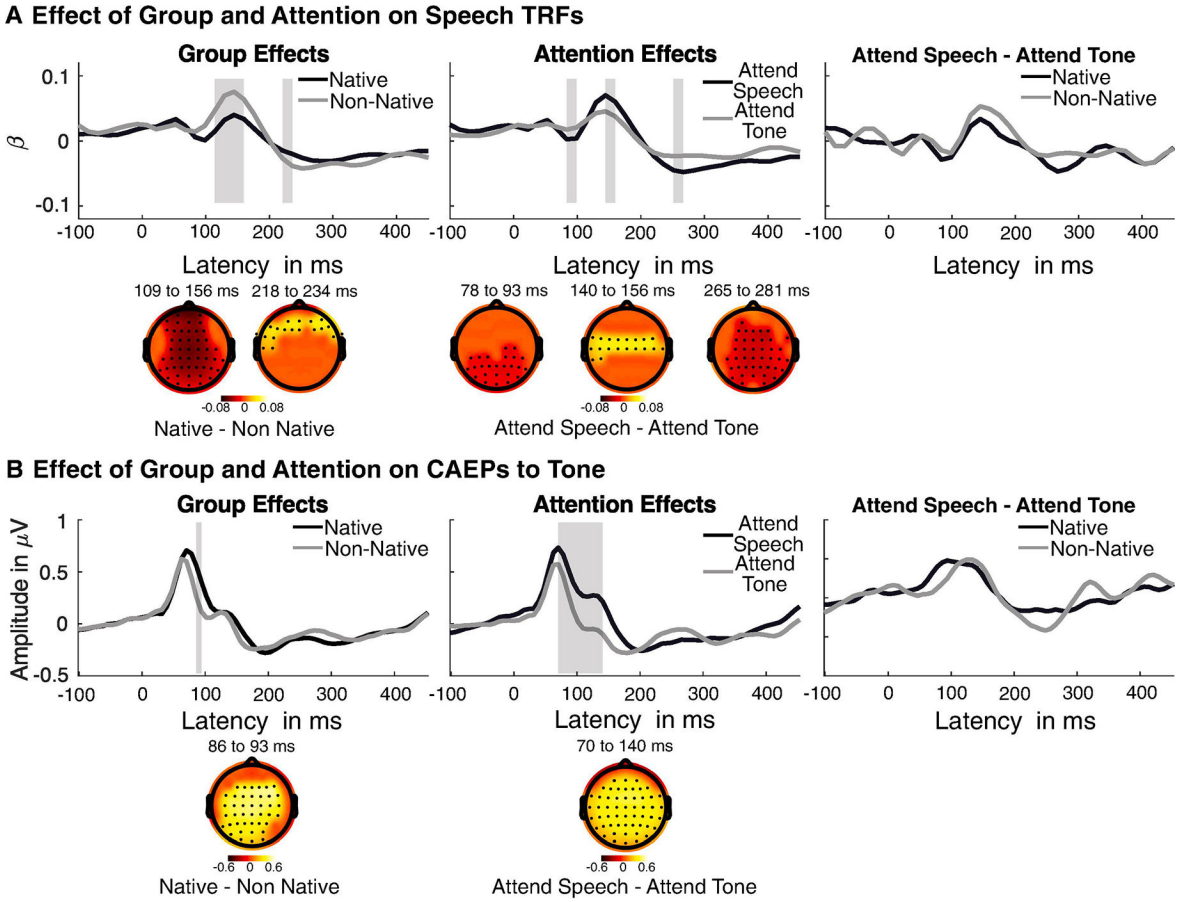

Figure 5. Results of point-wise comparison of the time courses of (A) TRFs to speech and (B) CAEPs to the tones.

Linear mixed effects analysis was performed to assess the effects of group and condition on the TRFs and CAEPs, and corrected for multiple comparison cluster-based permutation. Waveforms shown are averaged across attention conditions to show group effects (left panels), and averaged across groups to show attention effects (middle panel). Waveforms shown are averaged across channels that show significant effects. Highlighted regions shows latencies at which the waveforms show significant main effects. Differences in topography of the main effects are shown below each waveform for the different significant clusters. The topographic plots are masked to show electrodes with significant main effects. Extreme right panels show comparison between the groups for the attention effects (attend speech/tone - attend tone/speech), which did not show any significant group effects. Note: For the group and attention effects, the waveforms are averaged across the factors. While for the comparison of the attentional components in the two group, the absolute difference is shown.