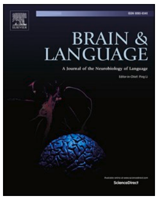
Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Continuous speech tracking in bilinguals reflects adaptation to both language and noise

Benjamin D. Zinszer [a], Qiming Yuan [b], Zhaoqi Zhang [b], Bharath Chandrasekaran [c], Taomei Guo [b,*]

[a] Department of Psychology, Swarthmore College, Swarthmore, PA, USA
[b] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, China
[c] Department of Communication Sciences and Disorders, University of Pittsburgh, USA

ARTICLE INFO

Keywords:
Speech perception in noise
Continuous speech
mTRF
L2 proficiency

ABSTRACT

Listeners regularly comprehend continuous speech despite noisy conditions. Previous studies show that neural tracking of speech degrades under noise, predicts comprehension, and increases for non-native listeners. We test the hypothesis that listeners similarly increase tracking for both L2 and noisy L1 speech, after adjusting for comprehension. Twenty-four Chinese-English bilinguals underwent EEG while listening to one hour of an audiobook, mixed with three levels of noise, in Mandarin and English and answered comprehension questions. We estimated tracking of the speech envelope in EEG for each one-minute segment using the multivariate temporal response function (mTRF). Contrary to our prediction, L2 tracking was significantly lower than L1, while L1 tracking significantly increased with noise maskers without reducing comprehension. However, greater L2 proficiency was positively associated with greater L2 tracking. We discuss how studies of speech envelope tracking using noise and bilingualism might be reconciled through a focus on exerted rather than demanded effort.

1. Introduction

Understanding continuous speech under noisy conditions is central to daily activities, like participating in a conversation or watching television in a restaurant. These conditions where noise or other voices mask a target speech signal (speech perception in noise; hereafter, SPIN) amplify the cognitive demands of speech perception (Mattys et al., 2012). The demands of SPIN are even greater for non-native listeners, who have less experience with the language they are hearing and possible interference from knowledge of their first language (Meador et al., 2000). Behavioral SPIN experiments have reliably shown that second-language listeners have greater difficulty than their monolingual peers perceiving individual words and sentences under various noise masks (see Garcia Lecumberri et al., 2010 for a review), measured by lower accuracy rates in word identification or recall. While this behavioral approach to SPIN tests for significant changes in how accurately monolingual or bilingual listeners infer exact speech content, it does not capture differences in the difficulty of the less extreme listening conditions that do not yield significant changes in performance. That is, while it is common to repeatedly fail to recognize words or comprehend sentences in a SPIN task, in natural conversation, listeners work to sustain comprehension by continuously modulating their effort and using additional contextual information to overcome noise.

Theoretical models of SPIN describe an interactive system where higher level linguistic processes can be modulated to compensate for masked or incomplete acoustic signals (e.g., the ELU model; Rönnberg et al., 2013). Smith and Fogerty's (2017) error pattern analysis of transcriptions of masked sentences confirmed that native English listeners were much more likely to substitute entire syntactically- and semantically-plausible words than more phonologically similar words that better matched the audible fragments of speech in a masked sentence transcription task. Non-native (bilingual) listeners exhibit analogous patterns to monolinguals, appearing to draw on linguistic knowledge to infer masked speech, despite lower overall accuracy (Bradlow & Alexander, 2007; Zinszer et al., 2019). These findings reveal how important adaptive recruitment of resources is in SPIN performance, even for second-language listeners with less language-specific information to draw on.

* Corresponding author at: State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China.
E-mail address: guotm@bnu.edu.cn (T. Guo).

https://doi.org/10.1016/j.bandl.2022.105128
Received 8 September 2021; Received in revised form 13 April 2022; Accepted 21 April 2022
Available online 7 May 2022

Although it is clear that listeners adapt to task demands, listener effort remains an ill-defined and poorly understood construct in general. One simple definition for listener effort is the combination of demanded effort (e.g., the difficulty of the task, the degradation of the signal) and exerted effort (e.g., the motivation of the participant), two separate but related factors which are confounded when effort is measured by performance (Francis & Love, 2019). In addition to attributes of the SPIN task, many attributes of the listener–such as second-language status–may influence demanded effort and, consequently, the difficulty of speech perception (Mattys et al., 2012). Exerted effort is more difficult to measure and has often been evaluated through self-report (e.g., Dimitrijevic et al., 2019), but exerted effort may also be accessible by physiological measure. Enhanced cortical tracking of an auditory speech signal is one effect of greater recruitment of cognitive resources (Zoefel & VanRullen, 2015), which can provide a physiological index of listener effort (Francis & Love, 2019).

Neural tracking measures compare the neurophysiological signal (such as a continuous EEG response) with a model of some property of the stimulus. The envelope, roughly the intensity of the speech sound (or rectified Hilbert transform of the auditory speech signal) contains crucial information supporting perception of both phoneme and sentential level speech (Shannon et al., 1995). Previous studies have supported the view that neural tracking of the speech envelope reflects listening effort by using EEG or MEG to demonstrate that attended speech is more strongly represented in the neural signal than unattended speech (Ding & Simon 2012; Rimmele et al., 2015).

For non-native listeners, we might infer that when matched with native listeners at comparable levels of speech comprehension in a SPIN task, exerted effort is greater than for the native listeners. Previous research has supported this claim: In behavioral research, late-bilinguals (first language speakers of Spanish, Hindi, Korean, Japanese, and other languages) needed + 10 dB signal to noise ratio (SNR) to perform at the same level of accuracy as English monolinguals in the QuickSIN task (Bidelman & Dexter, 2015). Further, non-native listeners have shown stronger speech tracking relative to monolinguals when listening to both short and continuous masked speech stimuli. Song and Iverson (2018; 2019) observed that Korean-English bilinguals showed greater neural tracking of English stimuli in the delta-theta band than monolingual listeners, while still performing far worse in an anomalous sentence detection task. This finding is corroborated by Reetzke et al. (2021) who also found that Mandarin-English bilingual listeners tracked continuous, masked English speech more strongly than native, monolingual listeners, although their performance on comprehension questions was nearly native-like. Translating this observation into a within-subjects context, bilingual listeners likely exert greater effort listening to their second language than to their first language, and we might predict that they show stronger speech tracking in L2 as compared to L1, when signal-to-noise and comprehension are balanced across the two languages.

Although stronger tracking seems to be associated with greater effort for the non-native listener, it does not always correspond with task demand, such as greater noise, nor is tracking of the speech envelope always predictive of better speech comprehension. Some studies report that neural tracking is positively related to speech intelligibility (Peelle et al., 2013), while others find no such relationship (Tune et al., 2020). McHaney et al. (2021) found that individual differences in older adults' ability to modulate their delta-band tracking of continuous speech (i.e., an audiobook) correlated with individual differences in their performance on comprehension questions. This study contrasted each participant's EEG tracking of the speech envelope of an unmasked audiobook to tracking of the same audiobook in noise. Participants who increased tracking during masked speech (relative to unmasked speech) showed significantly better comprehension than participants whose tracking remained the same or decreased during masked speech, even after controlling for hearing acuity and working memory. This finding provides unique evidence that listeners' change in the cortical response to the speech envelope corresponds with a change in exerted effort under greater task demands.

The similarity of the responses observed in non-native listeners relative to native listeners and in native listeners under masked listening conditions suggests that the same mechanism may support their respective adaptations: Increases in listening effort to accommodate greater task demand. Bilingual listeners, like monolingual listeners, also increased tracking at lower signal-to-noise ratios, even as behavioral performance simultaneously decreased (Skoe, 2019). Tracking in this study was estimated for fundamental frequency rather than amplitude envelope, and other levels of linguistic representation, such as phonemes (Di Liberto et al., 2015) and semantic dissimiliarity (Broderick et al., 2018) can be identified in EEG tracking and discriminate between levels of second language proficiency (Di Liberto et al., 2021).

Nonetheless, the speech envelope provides a relatively transparent, low-level property of acoustic stimuli with localized cortical responses that correlate with behavioral performance. Localization data provides mixed evidence about whether bilinguals and monolinguals draw on the same neurocognitive mechanisms when modulating representation of an acoustic signal. Although monolinguals' speech tracking is typically driven by activity in and around the auditory cortices (Horton et al. 2013; Golumbic et al. 2013), behavioral performance in SPIN tasks is correlated with response in the inferior frontal gyrus (IFG; Bidelman & Dexter, 2015). For bilinguals, activity in the IFG was not a predictor of behavioral accuracy, but activity in the superior temporal gyrus was (Bidelman & Dexter, 2015). This dissociation suggests that monolingual and bilingual listeners may adapt to noise by engaging different processes, and while the effective bilingual response (i.e., positively associated with comprehension) appears to amplify activity in regions associated with auditory processing, the effective monolingual response recruits other (possibly higher-order) processes.

To date, speech in noise studies using continuous stimuli have studied native and non-native processing between monolingual and bilingual participants in the same language, rather than directly contrasting first and second languages within bilinguals. Although evidence suggests that non-native listeners track the speech envelope more closely to support L2 comprehension, it's not known whether this response is characteristic of bilingual status in general, or whether the bilingual's tracking of a comparable stimulus in their L1 decreases relative to the L2 stimulus. Further, if listeners' exerted effort is the common cause of increased tracking for both non-native listeners and for decreases in signal-to-noise ratio, then we hypothesize that they are each independently associated with increased tracking at the individual subject level as long as comprehension is held constant.

In this study we designed a single continuous listening experiment aimed at testing our overarching hypothesis that bilinguals' EEG-based tracking of the acoustic envelope of speech in their non-native language reflected an adaptive response like that of native listeners under adverse conditions, such as lower signal-to-noise ratio (SNR). Specifically, we tried to reconcile findings from previous studies of bilinguals with the wider literature on speech tracking by contrasting two different challenges to speech comprehension: non-native language and steady-state noise masks. We predicted that decreases in tracking previously observed for noise-masked and non-native language occurred when difficulty was too great to be overcome by adaptation, and therefore we also compared the effects of including track-level comprehension accuracy in the model.

Under our hypothesis, we made three main predictions that would disentangle the effects of language proficiency, noise, and comprehension between individual participants. These predictions were preregistered with the Open Science Framework (OSF) prior to initiating the study: (1) In listeners' first language (L1), tracking of the EEG signal to the speech envelope increases under some noise (5 dB signal-to-noise ratio) but decreases under a more severe noise condition (0 dB signal-to-noise ratio) when not controlling for response accuracy. However, when data in the model are weighted by response accuracy, a consistent positive relationship exists between noise and speech tracking. (2) In listeners' second language (L2), speech tracking decreases under all noise conditions before adjusting for accuracy, but a positive relationship exists between noise and speech tracking when the model is accuracy-weighted. (3) Proficiency in L2 is associated with decreased tracking in participants and conditions, when adjusting for listening comprehension. Thus, tracking decreases with higher proficiency.

2. Method

In this experiment, Chinese-English bilingual participants listened to the first hour of an audiobook, *The Old Man and the Sea*, in two separate sessions. Across the two sessions we manipulated language, so that participants heard half part of the book in American English and the other half part in Mandarin Chinese on separate days. Within each session, we manipulated the signal-to-noise ratio (SNR) of the audiobook with three levels of speech-shaped, steady-state noise: No noise mask (just the audiobook), a 5 dB SNR condition (some noise), and a 0 dB SNR condition (more noise than 5 dB SNR). This design allowed us to contrast two forms of adverse listening conditions for speech comprehension (non-native language and noise mask) and test the predictions described in the foregoing introduction. Lastly, we evaluated the success of listeners' speech comprehension after each minute-long track with two multiple-choice comprehension questions administered in the same language as the auditory stimulus.

We pre-registered the experimental methods, hypotheses, and analysis plan on the Open Science Framework prior to beginning data collection, viewable at this link: https://osf.io/xyqfr/.

2.1. Participants

We recruited 27 Chinese-English bilinguals living in Beijing, China using posts on electronic bulletin boards (BBS) and other social media. Three participants dropped out in the middle of the experiment and the remaining 24 participants (11 male, 13 female) were retained for analysis. Their mean age was 22.0 years (SD = 2.7, range: 19–30). All participants were native speakers of Mandarin Chinese and did not report any cognitive, hearing, or other neurological disabilities. We aimed to recruit participants with generally moderate but varying levels of English language proficiency, measured by language production tasks and standardized test scores (College English Test Band 4, CET-4). We excluded any participants who had lived or traveled in an English-speaking country for greater than one month.

2.2. Materials

CET-4 scores. Previous completion of the CET-4 exam was a requirement for participation in this study. The CET-4 is often required for an undergraduate degree in China, and most prospective participants would have completed the exam recently. Scores range from below 330 (the first percentile) to above 650 (the 99th percentile), with median performance around 500. Although we are not aware of any psycholinguistic validity testing of the CET-4, it is designed to require an English vocabulary of approximately 4500 words, and scoring includes a wide range of English language skills (listening, reading, skimming, writing, and translation). The reported scores are based on the standardized distribution of test-takers' raw scores, thus providing a relatively straightforward metric of performance relative to peers. Further, the CET-4 is widely accepted in China for college and employment evaluation purposes. For further details, see the National College English Test Band 4 and Band 6 (2011) norm tables.

English verbal fluency task. Participants completed two verbal fluency tasks in English, one semantic and one phonemic. Semantic fluency is measured as the number of items the participant names (in English) from the categories clothing, furniture, animals, and fruits within one minute (per category). Phonemic fluency is measured as the number of words the participant names that begin with the letters F, A, and S within one minute (per initial letter). Participants' responses were audio recorded and transcribed by two Mandarin-English bilingual researchers. Scores for semantic and phonemic verbal fluency are the raw counts of items generated across categories for each task.

Mandarin verbal fluency task. We also tested Mandarin semantic fluency in the bilingual participants to provide a performance baseline in their native language. This task is not used in the estimate of English proficiency but is useful to identify outlier participants independent of the performance in the English tasks. As in English, participants were asked to name as many items as they could from the categories clothing, furniture, animals, and fruits within one minute per category in Mandarin Chinese. Altogether the English and Mandarin verbal fluency instructions and tasks took about 15 min to complete.

Audiobook. The audiobook stimuli were derived from two recordings of *The Old Man and the Sea* by Ernest Hemingway, one read by a male native speaker of Mandarin Chinese and the other by a male native speaker of American English. Although the original recordings are not licensed for free distribution (copyright owned by the original publisher), we have provided the speech envelope data from each recording necessary to estimate the mTRF model.

These versions were selected because the readers' voices and reading speeds were similar, allowing both versions of the book to be cut into segments that covered the same portions of the text and were each approximately the same duration. We estimated pitch for the tracks in each audiobook using Praat. Mean pitch in the English audiobook was 103 Hz, and mean pitch in the Mandarin audiobook was 136 Hz. Average duration of the English tracks was 59.5 s, and average duration of the Mandarin tracks was 64.3 s. Periodograms of the two audiobooks are depicted in Fig. 1. For each track of the audiobooks, we estimated the acoustic envelope by taking the absolute value of the Hilbert transform of the original.wav files. We then resampled the envelope data from 44.1 kHz down to 128 Hz (see envelopify.m and make_envelope_arrays.m on the OSF.io repository for the code used to generate the envelope data).

Noise Masks. The auditory masks were built based on sixty tracks (or about sixty minutes) from each audiobook (English and Mandarin). We generated the masks by estimating the long-term spectral distribution across all 120 tracks and shaped white noise to match the spectral distribution. For each audiobook track, we created white noise of equal duration and adjusted the amplitudes to create a track for each of the noise conditions (5 dB SNR and 0 dB SNR) with an overall intensity held constant across all three conditions.

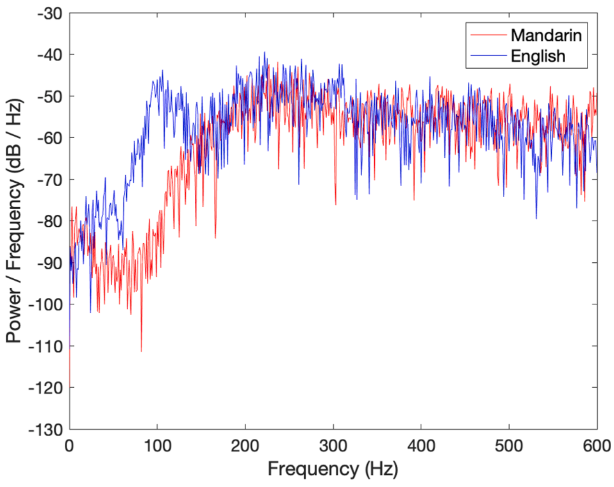


Fig. 1. Periodograms for the English audiobook (blue) and the Mandarin audiobook (red).

2.3. Procedures

Data collection. We piloted the stimulus presentation scripts and equipment in March 2019 (including two test participants whose scans were used to quality-check the processing pipeline and were not included in the analyzed dataset). The official recruitment of participants began in September 2019 after pre-registration of the method and hypotheses. Data collection was performed in November and December 2019, based on equipment availability.

In accordance with our OSF-registered plan, EEG data were quality checked at the halfway point (n = 12 participants) to verify that no major unexpected EEG artifacts appeared in the raw signal, mTRF correlations greater than zero were observed across most of the participants (regardless of presence or absence of between-condition differences), and to correct any procedural issues. None of these conditions for halting data collection were met, and we proceeded to collect the remaining 12 participants for a complete dataset.

Behavioral testing. After completing the informed consent, participants were asked to fill in a questionnaire including their language background and CET-4 scores. After that, participants were also asked to complete the verbal fluency task(s) in the language of the audiobook that they would later hear in that session.

EEG testing with audiobook. EEG data were collected on two separate days. Participants were asked to sit in a soundproof room where EEG data were recorded using a 64-channel Quik-Cap, a NEUROSCAN Synamps 2 amplifier, and the Acquire 4.2 software (NeuroScan Inc.). All electrodes were referenced online at an electrode placed between CZ and CPZ and were re-referenced offline to linked mastoids. Bipolar horizontal eye movement (HEOG) were recorded by 2 electrodes placed at the outer canthus of each eye, and bipolar vertical eye movement (VEOG) were recorded by 2 electrodes placed above and below the left eye. Both VEOG and HEOG were recorded for artifact rejection purposes. Electrode impedances were kept below 5 kilo-Ohms (kΩ). The sampling rate was 1000 Hz, and the online filtering was a bandpass between 0.05 and 100 Hz.

All participants heard the tracks in the same order: tracks 1 through 30 on the first day and tracks 31 to 60 on the second day. Both language and noise conditions were counterbalanced across these tracks according to the counterbalancing plan described on the OSF repository (https://osf.io/vpwc7/). Comprehension questions for each track were translation equivalents in English and Chinese, verified against the audiobook by native speakers of each language.

2.4. Analysis plan

Data analyses are replicable using the code provided on our OSF project page (https://osf.io/p49u6/). EEG preprocessing was performed using EEGLab in MATLAB, behavioral analyses were performed in R, and the models for hypothesis testing were estimated and evaluated in R. Details are provided in the sub-sections below.

English proficiency. For the purpose of testing our hypotheses about English proficiency, we summarize the verbal fluency data and CET-4 scores as a single index. We derive this measure by performing principal components analysis (PCA) over the CET-4 score, semantic fluency, and letter fluency, and we take the first component of this PCA as the English proficiency measure (as it accounts for the greatest amount of variance across subjects).

EEG preprocessing. EEG signal preprocessing was performed using the EEGLab package for MATLAB. Data were bandpass filtered between 1 Hz (high pass) and 15 Hz (low pass) and epoched in 70 s windows at each onset trigger, to cover the longest tracks. Excess data were trimmed later in analysis based on the duration of the acoustic data for each track. The beginning of each track was adjusted for the 500 ms delay between trigger onset and audiotrack onset. Data were resampled from 1000 Hz down to 128 Hz, which is a typical sampling rate for estimating mTRF of an acoustic envelope model.

We performed an extended independent component analysis (ICA) to identify ocular, electrode, and other artifacts. The typical ICA is often restricted to identifying a small subset of artifacts (e.g., eye-movement artifacts only) and has not been used to identify other types of artifacts such as high impedance electrodes and noise artifacts (Zou, Nathan, & Jafari, 2016). Therefore, the extended-ICA can be adopted to identify a broader range of artifacts since these artifacts all have identical independent component (IC) features (Hu & Zhang, 2019). The ICA was applied to compute ICs and an artificial screening process was subsequently applied to classify these ICs into artifact-related or neural-related ICs (Zou et al., 2016). The ICA rejection standard was set following the guideline provided in Hu and Zhang (2019): (1) power concentrated only in the frontal electrodes was removed as ocular artifacts; (2) power constrained within a single electrode was removed as electrode-related artifacts; (3) discontinued power topography was removed as other noise artifacts. The ICA screening and rejecting process was performed by two trained researchers independently, and the final rejecting decision was set as the union set of the two independent decisions to reach the best SNR (signal-to-noise ratio) of the final dataset.

Estimation of multivariate temporal response function (mTRF). Per the mTRF package authors' recommendation (see README.txt; MTRF toolbox, Crosse, Di Liberto, Bednar, & Lalor, 2016), we standardized the EEG data on a by-channel basis on-the-fly during analysis (z-score). These data were then analyzed in two stages of modeling: (1) We used multivariate temporal response function (mTRF) modeling to quantify the neural tracking of the EEG data to the stimulus model in each track, and (2) We use linear mixed effects modeling to test the effects of each manipulated variable (Language, Noise) while controlling for covariates (language order, noise order, CET-4 score, English verbal fluency). We describe stage (1) below and stage (2) in the following section.

The mTRF model is an extension of ridge regression modeling that has been implemented for fitting a temporal response function (the estimated parameters) to describe the relationship between a stimulus signal and a physiological signal. In many studies, like the present one, the comparison of interest is between a measurement of the acoustic envelope of speech as the predictor and multiple EEG channels as the outcome variable (i.e., the forward model, see Cross, Di Liberto, Bednar, & Lalor, 2016; Di Liberto, O'Sullivan, & Lalor, 2015). The complete specification and code for implementing the mTRF model can be found here: https://sourceforge.net/projects/aespa/files/. (Version 1.5 was the current version when the project was initiated. The most up-to-date version of the toolbox is now available at: https://github.com/mickcrosse/mTRF-Toolbox).

Three parameters must be specified prior to running the forward mTRF model: the time window onset, the time window offset, and the ridge parameter (lambda). We use a time window in the range [-100, 450] ms, matched with the previous studies of speech envelope tracking to which we compare our findings (McHaney et al., 2021; Reetzke et al., 2021). The ridge parameter is searched across a range of 10^[-5, 5] with increments of 1. The ridge parameter was selected independently for each participant: After estimating the model at each value of the ridge parameter for each track, the mean Pearson's *r* was calculated for that value of the parameter. This mean is combined across both languages, all three noise levels, and all channels. The ridge parameter yielding the highest overall *r* value was retained as the best model for that participant. Notably, this approach maximizes for higher correlations in general, but it avoids biasing the model towards higher or lower estimates in any given condition, language, or channel relative to the others.

The mTRF is trained and tested using an n-fold or leave-one-out cross validation approach, where each of n-tracks is held out as the test item once. The remaining n-1 tracks are concatenated to estimate the temporal response function linking the stimulus model (predictor) with the EEG data (outcome). After estimation, the convolution of the estimated temporal response function and stimulus model for the left-out track are Pearson correlated against the observed EEG data for the left out track.

This correlation is the r value reported for tracking of the EEG data to the stimulus in that track. The process is repeated n times (where n is the number of tracks), once for each track. In the present design, we entered all 60 of a participant's tracks into the n-fold design to estimate a participant-specific, but condition-generic mTRF. Each track was held out as the test data and the remaining 59 tracks used as training data.

Planned sanity checks. We planned and registered two sanity-check analyses to validate that our manipulations had the intended effects on listening difficulty. In the first sanity check, we predicted that overall mean accuracy on the comprehension questions in Mandarin (L1) would exceed mean accuracy in English (L2), indicating that the English language tracks presented a measurable increase in difficulty for the participants relative to the Mandarin language tracks. Our second sanity check predicted that comprehension would decrease with increasing noise in each language (Mandarin and English: Norm > 5 dB > 0 dB), but that these factors would interact. We speculated that both 5 dB and 0 dB noise in English would prove especially difficult for the bilingual listeners and not significantly different from one another, while Mandarin would show progressive decrease in comprehension across conditions.

Planned comparisons to previous effects. This experiment was designed to build on previous findings in mTRF studies, and we drew on aspects of two previous studies. In one study of Chinese-English bilinguals and English monolinguals (Reetzke et al., 2021), bilinguals living in the United States showed stronger tracking of English speech than their English monolingual peers. Therefore, we aimed to test this effect within participants by comparing L1 and L2 in the Norm condition, predicting that tracking of L2 speech (English) would be greater than tracking of L1 (Mandarin) in our group of bilinguals.

We also predicted that the comprehension accuracy would be positively related to subject-level differences in speech tracking, analogous to the previous finding that older monolingual adults who showed intersubject variation in comprehension of speech in noise, which was predicted by their tracking of the speech stimulus (McHaney et al., 2021). We predicted that under adverse listening conditions, subject-level Pearson r values would be positively correlated with comprehension scores. If participants showed above-chance comprehension (across the subject-level averages) in the 5 dB condition, we predict that subject-level mean r in this condition will positively correlate with their mean accuracy on comprehension questions. If the 5 dB condition does not result in better-than-chance comprehension, we made this prediction about the Norm condition instead.

Planned comparisons for hypothesis testing. The correlation scores resulting from the mTRF analysis are Pearson r statistics (correlation coefficients). Pearson r is not normally distributed by definition, since it is bounded on both ends, between −1 and 1. To improve the suitability of our data for a linear model, we transform the r values using the Fisher r-to-z transformation (otherwise described as the hyperbolic arctangent). This transformation is not likely to have any effect on our results, since typical r scores range from 0 to<0.20, over which the Fisher transform is almost linear. However, for the sake of statistical rigor, we committed in advance to use this transformation in case we encountered r values with larger magnitudes, where the transformation has a noticeable effect.

We use linear mixed effects models (implemented in R's lmerTest package; Kuznetsova, Brockhoff, & Christensen, 2017) estimated with the bobyqa optimizer (Powell, 2009) with 1E6 (one million) iterations to test our hypotheses about the behavioral and mTRF data. Model syntax is specified in the preregistration and described below. After estimating the models, we used post-hoc comparisons of the variables for Language, Noise, Proficiency, and their interactions, corrected using Tukey's HSD test for multiple comparisons (Tukey, 1977) using the emmeans function with Kenward-Roger estimated degrees of freedom (Lenth, 2021).

To estimate the effects of language and noise on comprehension, we entered the track-level response accuracy into a linear mixed effects model that estimates the fixed effects of Language and Noise (Language: L1 vs. L2; Noise: none, 5 dB, 0 dB). We included random intercepts for TrackID (1–60), OrderID (mask orders 1 through 12, see counterbalancing sheet), and Subject and random slopes for Language over TrackID and Subject to account for variance introduced by these aspects of the experimental design:

$$\begin{aligned}\text{Accuracy} \sim\ & \text{Language*Noise} + (1 + \text{Language}|\text{TrackID}) + (1|\text{OrderID}) \\ & + (1 + \text{Langauge}|\text{Subject})\end{aligned}$$

We also explored more complete random effects structures, including random slopes for each of the fixed effect variables and interactions over Track, Order, and Subject. This approach is recommended by Barr et al. (2013) instead of testing only random intercepts. None of the other models containing additional random slopes converged, suggesting that the reported model was best suited to these data.

Once all of the track-level r scores have been estimated for each participant, these measures (60 per participant × 24 participants) are transformed using Fisher's r-to-z and entered into a linear mixed-effects regression model that also includes Proficiency (a continuous measure based on the analysis described in the English Proficiency section). In the accuracy-controlled analyses, observations in the regression were weighted by Accuracy (0, 0.5, or 1) on the comprehension questions. This weighting is meant to remove trials where participants did not comprehend the audiobook (weight = 0), reducing the confounding of comprehension with the independent variables (language and noise) to improve the estimate of changes in speech tracking that may correspond to the effort needed to achieve that comprehension.

$$\begin{aligned}\text{Tracking\_z} \sim\ & \text{Language*Noise*Proficiency} + (1|\text{TrackID}) + (1|\text{Day}) \\ & + (1|\text{OrderID}) + (1|\text{Subject}), \text{weights} \\ & = \text{Accuracy}\end{aligned}$$

As in the comprehension data, we explored more complex random effects structures, beginning with random slopes for each fixed effect and interaction. None of these models converged, necessitating to the intercepts only approach instead.

3. Results

All results can be recomputed and replicated using the data and analysis scripts provided in the OSF project: https://osf.io/xyqfr. In a few cases, we deviated from aspects of the planned analyses. These deviations and any effects of the changes on the reported results are reported in the Supplementary Materials.

3.1. Behavioral results

English Proficiency. Twenty-two of our 24 participants completed the CET-4. After testing, we discovered that the remaining two participants had reported their scores on the TEM-4 instead, a more difficult test administered to English majors. We followed the pre-registered protocol by retaining these participants' EEG data in the sample and used mean imputation to replace their CET-4 scores. Although the test score data are not missing-at-random (i.e., English majors would reasonably be expected to have above average scores for this sample), we find that the CET-4 scores did not strongly affect the English proficiency ratings (further described below), and therefore these missing data did not greatly impact the results.

English verbal fluency scores were obtained for all 24 participants. In phonemic fluency, participants named an average of 11.60 words per onset phoneme (SD = 2.87). In semantic fluency, participants named an average of 8.59 objects per category (SD = 3.00). This mean was about half the number named for the same task in Mandarin (17.86, SD = 3.11), a statistically significant difference (paired-$t(23) = 12.9$, $p < 0.001$).

We performed the principal components analysis with varimax rotation over the three English scores (CET-4, English phonemic fluency, and English semantic fluency) to create a single English proficiency index using the psych package in R (Revelle, 2019), first by excluding the missing data from the CET-4 scores. The first component of the PCA accounted for 48% of variance in the data and heavily depended on the verbal fluency scores, which each had standardized loadings of 0.85, while the loading for the CET-4 was below 0.005. Mean imputation for the two missing CET-4 scores did not change these loadings, besides raising CET-4 to a standardized loading of 0.01 (see Table 1). We concluded that the imputation of two CET-4 scores had negligible impact on the PCA and used the imputed data to estimate Proficiency scores for all participants.

Comprehension Questions. We estimated mean accuracy on the post-track comprehension questions in each condition, controlling for random effects of Track, Subject, and counterbalancing Order using the emmeans package in R (Lenth, 2021). Subject-level mean values are depicted in Fig. 2A. We compared accuracy across languages and conditions using a linear mixed effects regression with fixed effects for Language and Noise, random slopes for language over track and subject, and random intercepts of Track, Day, Order, and Subject (package lmerTest for R; Kuznetsova, Brockhoff, & Christensen, 2017). The initial model was singular because Track and Day are correlated: Participants always heard tracks 1–30 on the first day and tracks 31–60 on the second day. We removed the random intercept of Day, re-estimated the model, and performed an ANOVA over the fixed effects. The interaction between Language and Noise condition was statistically significant (F (2,1273) = 12.190, $p < 0.001$).

We performed follow-up analyses of comprehension scores to determine whether our manipulations had the intended effects on listening difficulty. In the first comparison, overall mean accuracy in Mandarin (0.79) significantly exceeded English (0.50; t(40.4) = 11.34, $p < 0.001$), supporting our inference that comprehension of the L2 tracks was more difficult than the L1 tracks. In the second comparison, the noise manipulation (pooled across languages) was also effective in eliciting differences in comprehension. Accuracy in the Norm condition exceeded 5 dB SNR condition by 0.09 (t(1272) = 4.84, Tukey $p < 0.001$), 5 dB SNR exceeded 0 dB SNR by 0.06 (t(1272) = 3.42, Tukey $p = 0.002$).

Our final analysis of the comprehension data examined the individual noise conditions within each language, where we observed the significant interaction. Post-hoc tests were performed for all pairwise comparisons of the six Language × Noise conditions using Tukey's HSD correction. We predicted that comprehension would decrease with increasing noise in both Mandarin and English (Norm > 5 dB > 0 dB), but that English 5 dB and 0 dB might not significantly differ from one another if performance was especially low. Within the Mandarin conditions, none of the comprehension scores significantly differed between noise levels (Norm vs. 5 dB: $p = 1.00$; Norm vs. 0 dB: $p = 0.07$; 5 dB vs. 0 dB: $p = 0.21$), but all three Mandarin conditions significantly exceeded all three English conditions ($p \leq 0.005$). Further, within English, the Norm condition significantly exceeded both noise conditions ($p < 0.001$), but consistent with our speculation, English 5 dB and English 0 dB did not significantly differ ($p = 0.10$). Subject-level average comprehension scores for each condition are depicted as individual data points in Fig. 2A, with group-level averages depicted by the bar plots.

Table 1
Language proficiency scores.

Measure	Mean	SD	Range	Load on index
Mandarin				
Semantic fluency	17.86	3.11	13.5, 26.5	–
English				
Semantic fluency	8.59	3.00	3.0, 16.5	0.85
Phonemic fluency	11.60	2.87	5.3, 17.0	0.85
CET-4	570	34	487, 633	0.01
Proficiency Index	0	1	−1.63, 2.27	–

3.2. EEG results

mTRF Correlations. Mean subject-level Fisher r-to-z transformed mTRF correlations are depicted for each condition in Fig. 2B, with group-level averages depicted by the bar plots. Correlations from each track were entered as observations into the mixed effects linear models. One model was weighted by response accuracy, such that tracks with two correct responses were weighted more heavily than tracks with one correct response, and tracks with no correct responses were not included. The other model was unweighted such that all tracks were included regardless of accuracy on the comprehension questions. Weighting was performed using the weights argument in lmer function.

Both models included fixed effects of Language, Noise, and L2 (English) Proficiency and random intercepts for TrackID (1–60), Day (1–2), OrderID (1–12), and Subject (1–24). The unweighted model was singular. OrderID was removed from that model, which then converged. In the case of any significant effects (interactions or main effects), we used the emmeans function to make all possible pairwise comparisons between conditions, with Tukey's HSD adjustment for multiple comparisons. Contrary to our predictions, we did not find that the results of these two models (weighted vs. unweighted) greatly differed. Hereafter, we report both models where possible (see Table 2), but results are generally not affected by the choice of one model or the other.

Planned comparisons to previous effects. Following the finding of Reetzke et al. (2021), we compared the tracking of L1 and L2 in the Norm condition. In both the weighted and unweighted models, tracking of the Mandarin Norm condition significantly exceeded the English Norm condition (unweighted: difference = 0.017, t(1347) = 13.38, $p < 0.001$; weighted: difference = 0.016, t(427) = 12.26, $p < 0.001$; see Table 2).

Following the finding of McHaney et al. (2021), we correlated the r-to-z transformed tracking with mean comprehension accuracy scores in the 5 dB condition at participant-level. This correlation was moderate and statistically significant ($r = 0.39$, $p = 0.007$), but separating the correlations out by language, we found that this relationship was primarily driven by between language differences (English 5 dB tracks: $r = 0.04$, $p = 0.85$; Mandarin 5 dB tracks: $r = 0.06$, $p = 0.76$).

Planned hypothesis tests. The first hypothesis predicted that tracking of L1 would increase under noise (5 dB SNR condition) and then decrease when the noise increased (0 dB SNR) in the unweighted model, but it would instead increase under 0 dB SNR in the accuracy-weighted model. In both models, tracking of L1 did significantly increase from Norm to 5 dB (unweighted: difference = -0.017, t(1347) = -7.47, $p < 0.001$; weighted: difference = -0.017, t(888) = -8.91, $p < 0.001$). Tracking remained significantly elevated in the Mandarin 0 dB condition relative to Norm (unweighted: difference = -0.013, t(1347) = -5.87, $p < 0.001$; weighted: difference = -0.013, t(812) = -6.73, $p < 0.001$), unchanged from Mandarin 5 dB to 0 dB in both models (unweighted: difference = 0.004, t(1347) = 1.61, $p = 0.59$; weighted: difference = 0.004, t(803) = 2.00, $p = 0.34$).

The second hypothesis predicted that tracking of L2 would decrease for both 5 dB and 0 dB conditions in the unweighted model and increase for both conditions in the weighted model. Instead, we did not find statistically significant differences in tracking between English Norm and English 5 dB (unweighted: difference = -0.006, t(1347) = -2.67, $p = 0.08$; weighted: difference = -0.005, t(364) = -2.23, $p = 0.23$). From English 5 dB to 0 dB, both models reflected a significant decrease (unweighted: difference = 0.009, t(1347) = 3.97, $p = 0.001$; weighted: difference = 0.009, t(236) = 3.17, $p = 0.02$). The Norm and 0 dB conditions did not significantly differ in English (unweighted: $p = 0.78$, weighted: $p = 0.81$).

Last, we examined how individual differences in L2 Proficiency affected tracking of the English tracks. Based on previous comparisons of native to non-native English listeners, we predicted that an accuracy-weighted model would find decreased tracking in participants with greater proficiency. We re-estimated the linear mixed-effects model

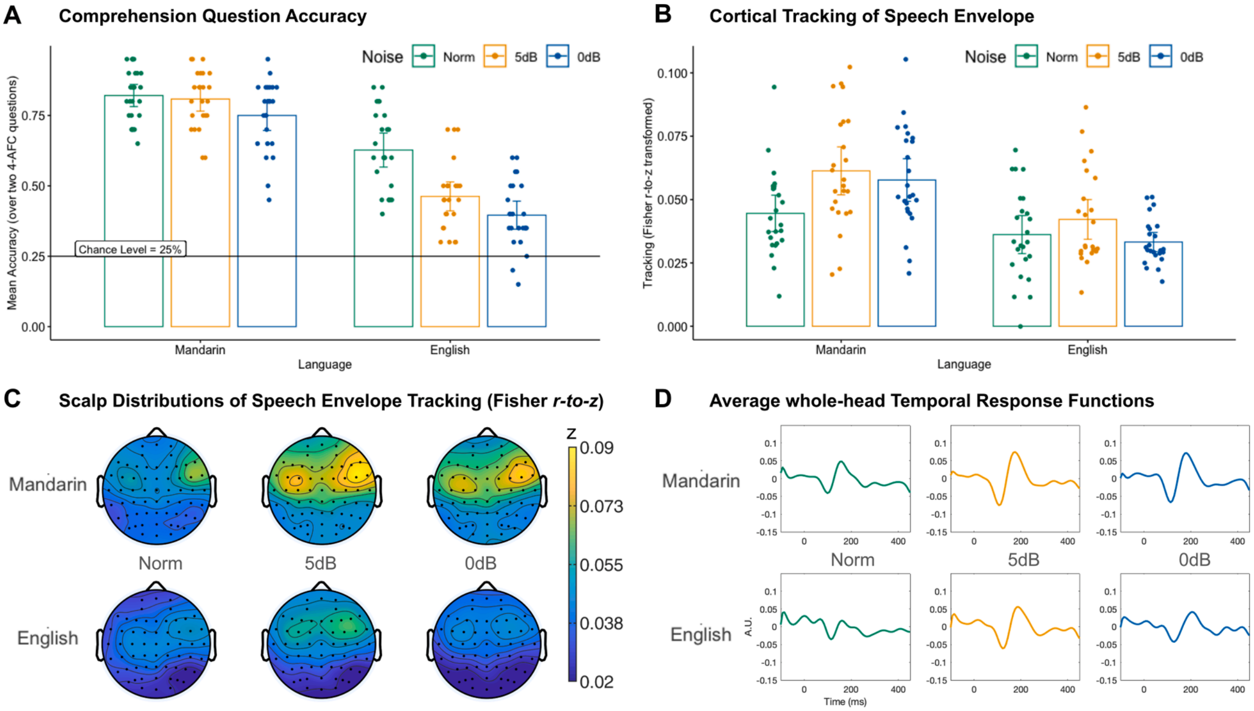


Fig. 2. (A) Mean subject-level accuracy on comprehension questions in each condition. Dots indicate individual subject-level results, error bars 95% confidence interval of the mean. (B) Mean EEG tracking of speech envelope in each condition. Dots indicate individual subject-level results, error bars 95% confidence interval of the mean. (C) Scalp distributions for EEG tracking in each of the six conditions, using Fisher's r-to-z transformed correlations at each channel. (D) Average temporal response function (TRF) for all channels (whole head) in each condition.

Table 2
Estimated marginal means for EEG tracking of the speech envelope.

	Accuracy-weighted model	Unweighted model
Mandarin Norm > English Norm	0.016, t(427) = 12.26, p < 0.001	0.017, t(1347) = 13.38, p < 0.001
Mandarin		
Norm – 5 dB	−0.017, t(888) = -8.91, p < 0.001	−0.017, t(1347) = -7.47, p < 0.001
Norm – 0 dB	−0.013, t(812) = -6.73, p < 0.001	−0.013, t(1347) = -5.87, p < 0.001
5 dB – 0 dB	0.004, t(803) = 2.00, p = 0.34	0.004, t(1347) = 1.61, p = 0.59
English		
Norm – 5 dB	−0.005, t(364) = -2.23, p = 0.23	−0.006, t(1347) = -2.67, p = 0.08
Norm – 0 dB	0.003, t(287) = 1.25, p = 0.81	0.003, t(1347) = 1.30, p = 0.78
5 dB – 0 dB	0.009, t(236) = 3.17, p = 0.02	0.009, t(1347) = 3.97, p = 0.001

using only English tracks, with fixed effects of L2 Proficiency, Noise, and their interaction and random intercepts for Track, Day, OrderID, and Subject. The interaction between Proficiency and Noise was statistically significant ($F(2,676.44) = 10.23, p < 0.001$), warranting an examination of the simple slopes for Proficiency at each Noise level. The simple slopes analysis revealed a significant, positive effect of Proficiency in the Norm condition (0.007 increase in tracking for every 1 SD in Proficiency, $p = 0.03$; see Fig. 3 for subject-level averages) and no significant effects of Proficiency in the 5 dB or 0 dB conditions (0 dB: −0.004, $p = 0.21$; 5 dB: 0.003, $p = 0.31$).

Scalp distribution and ROI analyses. The planned analyses were conducted across the full array of 62 electrodes (excluding the eye movement channels and two mastoid references). In these whole-head analyses, the estimates for tracking on each channel are averaged to yield a single summary statistic, but individual channels can also be disaggregated to observe the topography of neural tracking scores across the scalp (pictured in Fig. 2C).

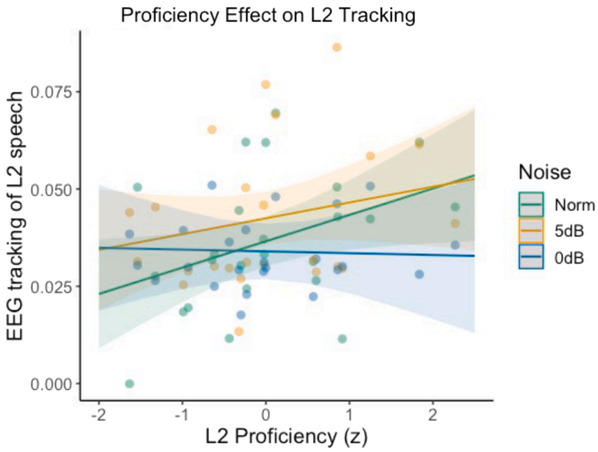


Fig. 3. Simple slopes for cortical EEG tracking to the speech envelope over second language proficiency. The positive slope for Norm condition was statistically significant, but the slopes in 5 dB and 0 dB conditions were not. Error fields are 95% confidence interval of the slope.

For a follow-up analysis of L2 proficiency effects, we defined nine regions of interest on the scalp based on three divisions of the electrodes on each axis of the horizontal plane: Anterior (AF, F, and FP electrodes), Central (C, T, FT, FC, CP, and TP electrodes), or Posterior (P, O, PO, and CB electrodes) and Left (odd numbered electrodes), Midline (Z electrodes), or Right (even numbered electrodes). Channel-level estimates were averaged within each ROI, and we re-estimated the accuracy-weighted model reported above with Noise, L2 Proficiency, and ROI as fixed effects and Subject, Track, Day, and OrderID as random intercepts.

The three-way interaction between Noise, ROI, and L2 Proficiency was not significant in the analysis of variance for this model (F(16, 6284.6) = 0.83, p = 0.65), but ROI significantly interacted with Proficiency (F(8, 6284.6) = 2.20, p = 0.024) and Noise (F(16, 6284.6) = 4.06, p < 0.001), as well as having a significant main effect (F(8, 6284.6) = 62.22, p < 0.001). Proficiency and Noise also significantly interacted (F (2,6385.1) = 46.41, p < 0.001). Noise had a significant main effect (F (2,6370.2) = 33.73, p < 0.001) while Proficiency did not (F(1,29.8) = 0.27, p = 0.61). Thus the nine ROIs differed not only in their mean response amplitudes, but in their effects of L2 proficiency (ROI*Proficiency) and differences between Noise levels (ROI*Noise).

We followed-up these two-way interactions by testing the pairwise contrasts for Noise and simple slopes of Proficiency in each of the nine ROIs separately. Within each ROI, Noise comparisons were Dunn-Bonferroni corrected (0.05/3). No effects survived between-ROIs correction, so we report the within-ROI corrected results: In all six anterior and central regions, L2 tracking in the 5 dB condition exceeded 0 dB (p < 0.05 corrected; individual p-values can obtained from the shared scalp_distribution_analyses code). In the three anterior regions and left central, tracking in 5 dB also exceeded tracking in the Norm condition (p < 0.05 corrected). The left posterior region also showed a significant difference between Norm and 0 dB (p = 0.011). See Fig. 4 for summary.

The simple-slopes analysis across ROIs indicated that the effect of L2 Proficiency on English speech tracking was statistically significant across the three central regions: left (0.008, p = 0.006), midline (0.007, p = 0.013) and right (0.007, p = 0.028). Further, the same effect was observed in the right anterior region (0.007, p = 0.023). See Fig. 4 for slopes in each ROI.

4. Discussion

In this study, we examined the relationship between listening effort and neural tracking of the acoustic envelope of natural, continuous speech in bilinguals. Previous studies suggested that tracking was stronger (i.e., a higher correlation between electrophysiological data and the stimulus) when the difficulty of the task was higher for bilinguals (i.e., for non-native listeners relative to native listeners; Reetzke et al., 2021; Song et al., 2019), but also that increases in tracking were positively associated with higher comprehension in monolingual listeners (Peelle et al., 2013; McHaney et al., 2021). How do we reconcile these two bodies of evidence? We proposed that listeners' exerted effort could explain these seemingly divergent accounts (consistent with the proposal of Francis & Love, 2019), and we predicted that after adjusting for comprehension, within-subject increases in speech tracking would be observed for both second language (L2) listening and increases in noise (i.e., decreased SNR). That is, these two distinctly different manipulations of the stimulus would each elicit greater listener effort and therefore stronger neural tracking of the speech envelope.

The present study directly manipulated both language (L1, Chinese and L2, English) and signal-to-noise ratio (No noise or Norm, 5 dB SNR, 0 dB SNR) within participants while they listened to approximately one hour of an audiobook. We estimated comprehension at sixty intervals (or about once per minute) using questions in the same language as the audiobook. On the group level, accuracy on the comprehension questions supported our basic sanity-checks that L2 listening was more difficult than L1 listening under identical SNR levels, and that increasing noise decreased accuracy in L2, but–crucial to understanding the results

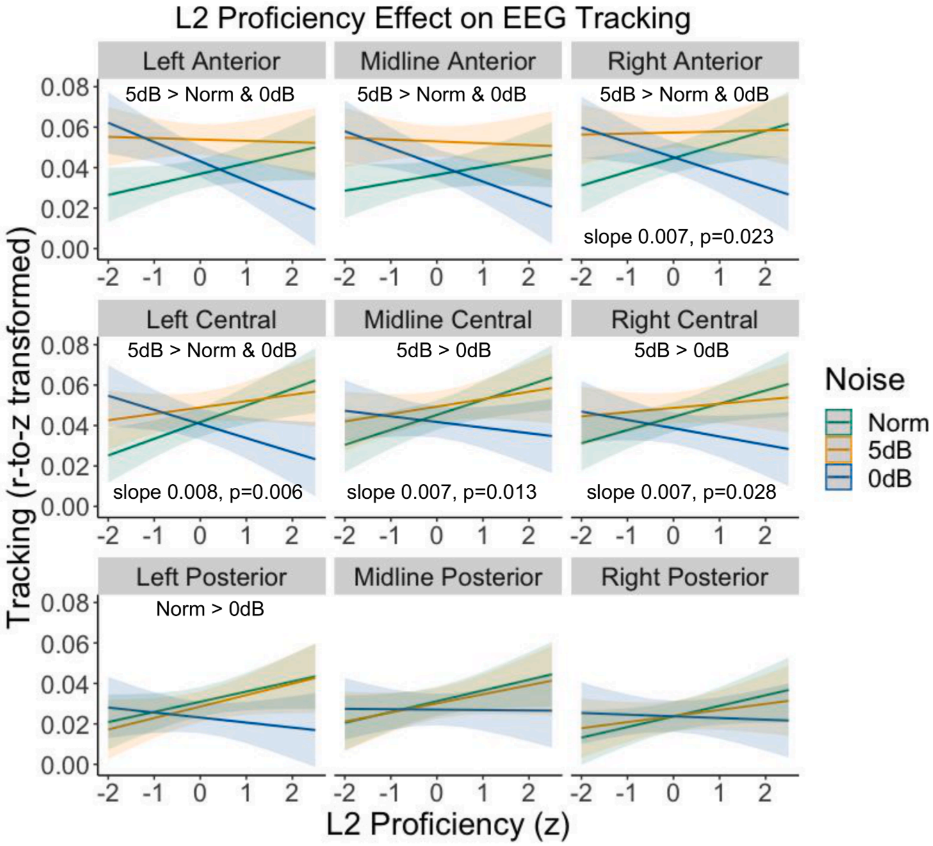


Fig. 4. Simple slopes interaction plot for effects of L2 proficiency and noise conditions on L2 tracking in the nine regions of interest. Error fields are 95% confidence interval of the slope. Text labels report pairwise contrasts of noise effect and overall L2 proficiency slopes in each region. No three-way interaction between L2 proficiency and Noise was observed, so individual noise-by-proficiency slopes are not compared to each other.

of the speech tracking analysis–comprehension did not significantly change in L1 between different noise conditions. However, comparison of the two models of cortical tracking, weighted by comprehension of individual tracks vs. not using the comprehension measure, did not suggest that this measure was important for understanding individual trial level variations in neural tracking.

4.1. Does L1 tracking increase with noise after adjusting for comprehension?

The close parity between the L1-Norm (no noise), L1-5 dB, and L1-0 dB comprehension provided a simple test case for our overarching prediction that listeners would increase neural tracking to the speech envelope in their native language (L1) to sustain comprehension under adverse listening conditions. Among the pairwise comparisons described in our analyses (see Fig. 2A), the incremental increases in L1 difficulty did not result in a significant decrease in comprehension, while the difference between no noise and 5 dB SNR in L1 did result in an increase in tracking that was sustained under greater noise (0 dB SNR, see Fig. 2B). We predicted that after adjusting for accuracy on comprehension questions, tracking would increase again for the 0 dB SNR condition relative to 5 dB SNR condition, but no such difference was observed. Therefore we did not find unequivocal support for our prediction that of a consistently positive relationship between speech envelope tracking and noise, with or without adjustment for performance on the comprehension questions.

4.2. Does L2 tracking increase with noise after adjusting for comprehension?

Unlike the pattern observed in L1, estimated tracking of EEG data to the second language (L2) did not significantly increase under noise, alongside lower overall comprehension in the L2 noise conditions relative to the unmasked L2-norm condition. This conclusion was not affected by weighting the data by accuracy on the comprehension questions. In both the weighted and unweighted models, L2-norm and L2-5 dB conditions did not significantly differ (although tracking in 5 dB was slightly numerically higher than in norm), but the L2-0 dB tracking decreased slightly (and statistically significantly), relative to L2-5 dB. This small change in the result might be taken to suggest that tracking and comprehension are still associated in the L2 noise conditions, but the considerable overall decrease in comprehension between L2-norm and L2-5 dB was not effectively offset by a change in neural tracking. However, the accuracy-weighted model not only failed to clarify this relationship, but actually slightly reduced the estimated difference between tracking in the Norm and 5 dB conditions. Moreover, participants' average accuracy in even the easiest L2 condition was significantly lower than the most difficult L1 condition, casting some doubt on the level of overall comprehension of the L2 stimuli.

These results differ from Reetzke et al.'s (2021) study, where L2 listeners were roughly matched with native listeners on average comprehension prior to making the speech envelope tracking comparison. In any given condition, participants may be correctly guessing some answers to multiple choice questions based on "glimpsed" words (see Cooke, 2006), but particularly at the lower accuracy level observed in the present study, the relative contribution of glimpsed or guessed answers can be much higher, highlighting a limitation of this measure for comprehension. Besides recruiting bilingual participants with higher L2 proficiency for comparison, another approach to this problem would be tuning the L2-norm condition for higher comprehension a priori (e.g., slower reading speed, easier vocabulary), comparable to L1 levels, and then looking for changes with manipulated signal-to-noise ratios.

4.3. Is L2 proficiency inversely related to tracking of the L2 speech envelope?

Our within-subjects comparison of L1 and L2 in bilingual listeners diverges from previous between-subjects studies of native vs. non-native listeners (Reetzke et al., 2021; Song et al., 2018, 2019), finding that bilingual listeners tracked their L2 more weakly than their L1. However, one crucial difference between these two approaches (between- vs. within-subjects) is that the non-native listeners in the between-subjects studies were immersed in an L2 environment and therefore had relatively high L2 proficiency. In the present study, our bilingual listeners were still immersed in their native language environment and had more moderate L2 proficiency, as evidenced by their median to moderately-high CET-4 scores, their lower performance on L2 fluency tasks compared with the L1 fluency task, and lower performance on the L2 comprehension questions than L1 across all conditions.

In light of that evidence, we consider whether individual differences (between subjects) in L2 proficiency predicts differences in tracking of the L2 speech envelope. Contrary to our original predictions, L2 tracking significantly increased with L2 proficiency in the easiest (no noise) listening condition. When broken down by ROI, significant positive relationships between L2 proficiency and L2 tracking were observed across left, midline, and right central regions and the right anterior region. Although the lack of a region-by-noise-by-proficiency three-way interaction did not permit comparison of these proficiency effects between different noise conditions, the positive relationship appeared to be driven by the norm and, to a lesser extent, the 5 dB conditions.

If this pattern holds across a range of higher proficiency, L2-immersed bilinguals, it would indicate a crucial difference between languages in bilingual listeners: As listeners approach native-like proficiency levels in L2, the differences between L1 and L2 cortical tracking would widen on the whole-head level. However, it is also possible that, in this study, we have observed the left side of an inverted U-shaped function. If the U-shaped prediction is correct, then tracking to L2 would increase with L2 proficiency (as observed in this study) up to a certain threshold at which L2 comprehension can match L1 comprehension behaviorally with greater listening effort (reflected in envelope tracking, as in Reetzke et al., 2021), and beyond that inflection point, effort and tracking decline with proficiency while ceiling-level comprehension is sustained. Only additional experiments will clarify between these alternative explanations.

4.4. What do changes in neural tracking tell us about bilinguals' adaptation in L1 and L2?

Although the present findings differed considerably from our preregistered predictions, we see an emerging pattern among the most recent literature on listening effort in native speakers and on comparing native and non-native listeners. Critically, listening effort has been measured differently between studies, sometimes measured with self-report (e.g., Dimitrijevic et al., 2019), sometimes with manipulations of signal quality or signal-to-noise ratio (e.g., Hauswald et al., 2020; the present study), and sometimes based on between-subject differences in task difficulty (e.g., Song et al., 2018). Each of these measures bears on ultimate listener effort, but by measuring different contributors: either demanded effort or exerted effort (Francis & Love, 2019). It remains true that demanded effort and exerted effort should be positively correlated (or confounded), so long as performance is constant, but performance has rarely been controlled as a covariate in the relationship between task and neural tracking.

In conjunction with other studies, a picture is emerging that adaptations for listeners at different levels of language proficiency and in hearing different levels of signal to noise might be analogous. Dimitrijevic et al. (2019) controlled demanded effort in participants with cochlear implants by setting SNR relative at their comprehension threshold prior to the experiment. In their between-subjects comparison, delta band coherence in the left frontal cortex was negatively associated with reported listening effort, but in their within-subjects comparison, left temporal coherence was positively associated with performance on the task. Scalp distributions in our data suggested a right-lateralized response to noise, with the strongest tracking over right anterior or central regions. However, L2 proficiency effects were also bilateral and central, perhaps suggesting a common mechanism with Bidelman and Dexter's (2015) bilinguals completing a speech in noise task in their L2.

A within-subject account of adaptively increasing neural tracking to the speech envelope to compensate for signal degradation is also consistent with McHaney et al.'s (2021) finding that greater upregulation of tracking was associated with better comprehension in a between-subjects comparison of monolingual older adults. Likewise, our study finds stronger tracking of L1 under noise while comprehension remains unchanged. Both of these responses, modulation in response to masking of L1 and in response to L2 status, were qualitatively similar in their right-lateralized frontocentral scalp distributions, possibly pointing to a unifying account of listening effort that underlies both sets of responses.

On the other hand, the spatially localized effects on L2 tracking were not the same for between-subjects differences in L2 proficiency (bilateral and central) and within-subjects differences in mask intensity (anterior). This divergence suggests that different mechanisms may underlie these adaptations of neural tracking to the speech envelope for different sources of demanded effort. Studies that have estimated tracking for linguistic variables other than the speech envelope have found an important role of bilingual status in cross-language phonemic differences (Di Liberto et al, 2021) and fundamental frequency (Skoe, 2019). The diversity of linguistic cues available to monolingual and bilingual listeners may help to account for differences between these groups in which regions predict behavioral performance (Bidelman & Dexter, 2015).

5. Conclusion

Overall, the results of both the L1 and L2 noise manipulations challenge the simplicity of a monotonic relationship between demanded effort (such as noise or bilingual status) and neural tracking of the acoustic envelope for continuous speech. The current results concur with emerging patterns from both native and non-native listeners that neural tracking may instead reflect exerted effort when outcomes (i.e., comprehension) are held constant. Our findings diverge from previous studies with higher proficiency bilinguals, but the observed positive relationship between L2 proficiency and tracking suggests that the present study and previous studies may not be irreconcilable: Further tests should include more targeted manipulations of demanded effort, such as L1/L2 status and SNR, alongside subjective factors like L2 proficiency, while maintaining comprehension or controlling for it with behavioral measures (which the present study achieved only coarsely). We predict that under an exerted-effort account, more refined studies will still find that L2 neural tracking increases with L2 proficiency to the point of exceeding L1 tracking when overall behavioral performance between L1 and L2 are matched, and L1 tracking will increase with greater noise as long as comprehension remains constant.

Competing interests: The authors declare none.

Declaration of Competing Interest

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.

Data availability

All results can be recomputed and replicated using the data and analysis scripts provided in the OSF project: https://osf.io/xyqfr

Acknowledgements

The study was supported by the Open Grant of State Key Laboratory of Cognitive Neuroscience and Learning to Benjamin D. Zinszer, and the National Natural Science Foundation of China (31871097) to Taomei Guo. We are grateful to Sarah Campbell, Priyanka Deshmane, and Jacie McHaney for their assistance with stimulus preparation, Rachel Reetzke and Heather Dial for their consultation in planning the project, to Yingying Sun, Qianwen Chang, and Zeping Liu for their effort in data collection, and to Man Zhang for her generous help in analyzing EEG data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2022.105128.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bidelman, G. M., & 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. https://doi.org/10.1016/j.bandl.2015.02.002

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology, 28*(5), 803–809. https://doi.org/10.1016/j.cub.2018.01.080

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience, 10*, 604. https://doi.org/10.3389/fnhum.2016.00604

Di Liberto, G. M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. A., & Mesgarani, N. (2021). Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage, 227*, Article 117586. https://doi.org/10.1016/j.neuroimage.2020.117586

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology, 25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports, 9*(1), 1–10. https://doi.org/10.1038/s41598-019-47643-1

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Francis & Love. (2019). WIREs Cognitive Science. https://doi.org/10.1002/wcs.1514.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Hauswald, A., Keitel, A., Chen, Y. P., Roesch, S., & Weisz, N. (2020). Degradation levels of continuous speech affect neural speech tracking and alpha power differently. *European Journal of Neuroscience, 1–15*. https://doi.org/10.1111/ejn.14912

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology, 109*(12), 3082–3093. https://doi.org/10.1152/jn.01026.2012

Hu, L., & Zhang, Z. (Eds.). (2019). *EEG Signal Processing and Feature Extraction*. Springer Singapore.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lenth, R. V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0. https://CRAN.R-project.org/package=emmeans.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McHaney, J. R., Gnanateja, G. N., Smayda, K. E., Zinszer, B. D., & Chandrasekaran, B. (2021). Cortical tracking of speech in delta band relates to individual differences in speech in noise comprehension in older adults. *Ear and Hearing, 42*(2), 343–354. https://doi.org/10.1097/AUD.0000000000000923

National College English Test Band 4 and Band 6. (2011). Retrieved by *Internet Archive* on March 23, 2019. https://web.archive.org/web/20190323181532/http://www.cet.edu.cn/cet2011.htm.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex, 23*(6), 1378–1387.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Department of Applied Mathematics and Theoretical Physics, Cambridge England, Technical Report NA2009/06*.

Reetzke, R., Gnanateja, G. N., & Chandrasekaran, B. (2021). Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain and Language, 213*, Article 104891. https://doi.org/10.1016/j.bandl.2020.104891

Rimmele, J. M., Golumbic, E. Z., 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. https://doi.org/10.1016/j.cortex.2014.12.014

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., … Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7*, 31. https://doi.org/10.3389/fnsys.2013.00031

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303–304.

Skoe, E. (2019). Turn up the volume: Speech perception in noise for bilingual listeners. *The Journal of the Acoustical Society of America, 145*(1820). https://doi.org/10.1121/1.5101649

Smith, K. G., & Fogerty, D. (2017). Speech recognition error patterns for steady-state noise and interrupted speech. EL306-EL312 *The Journal of the Acoustical Society of America, 142*(3). https://doi.org/10.1121/1.5003916.

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. https://doi.org/10.1016/j.cognition.2018.06.001

Song, J., Martin, L., & Iverson, P. (2019). Native and non-native speech recognition in noise: Neural measures of auditory and lexical processing. *International Congress of Phonetic Sciences*.

Tukey, J. W. (1977). *Exploratory data analysis, 2*, 131–160.

Tune, Alavash, Fiedler, & Oblese. (2020) "Neural attention filters do not predict behavioral success in a large cohort of aging listeners" https://www.biorxiv.org/content/10.1101/2020.05.20.105874v1.full.pdf.

Zinszer, B. D., Riggs, M., Reetzke, R., & Chandrasekaran, B. (2019). Error patterns of native and non-native listeners' perception of speech in noise. *The Journal of the Acoustical Society of America, 145*(2EL129-EL135). https://doi.org/10.1121/1.5087271

Zoefel, B., & VanRullen, R. (2015). The role of high-level processes for oscillatory phase entrainment to speech sound. *Frontiers in Human Neuroscience, 9*, 651. https://doi.org/10.3389/fnhum.2015.00651

Zou, Y., Nathan, V., & Jafari, R. (2014). Automatic identification of artifact-related independent components for artifact removal in EEG recordings. *IEEE Journal of Biomedical and Health Informatics, 20*(1), 73–81. https://doi.org/10.1109/JBHI.2014.2370646