Neural representation of linguistic feature hierarchy reflects second-language proficiency

Giovanni M. Di Liberto[1,*], Jingping Nie[2,3,*], Jeremy Yeaton[1,4], Bahar Khalighinejad[2,3], Shihab A. Shamma[1,5], Nima Mesgarani[2,3,**]

1 Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France
2 Department of Electrical Engineering, Columbia University, New York, NY, USA
3 Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, United States
4 Laboratoire de Psychologie Cognitive, UMR 7290, CNRS, France. Aix-Marseille Université, France.
5 Institute for Systems Research, Electrical and Computer Engineering, University of Maryland, College Park, USA

** The authors contributed equally*
*** Senior author*

Correspondence: diliberg@tcd.ie, nima@ee.columbia.edu

Conflicts of interest: none declared.

Funding sources: this study and N.M., B.K., and J.N. were supported by the NIMH MH114166-01 grant. G.D.L. and S.S. were supported by an ERC Advanced grant (Neume, Grant agreement ID: 787836). J.Y. was supported by the ERC Advanced grant (POP-R, Grant agreement ID: 742141).

Word count (excluding abstract, title page, references and methods): 5798.

Acknowledgements: The authors would like to thank Michael Broderick for his help with the semantic dissimilarity analysis. The authors also thank Adam Soussana and Ghislain de Labbey for their help with a pilot version of this experiment.

Abstract

Acquiring a new language requires a simultaneous and gradual learning of multiple levels of linguistic attributes. Here, we investigated how this process changes the neural encoding of natural speech by assessing the encoding of the linguistic feature hierarchy in second-language listeners. Electroencephalography (EEG) signals were recorded from native English and native Mandarin speakers with varied English proficiency as they listened to English stories. The neural encoding of acoustic, phonemic, phonotactic, and semantic features was measured in individual participants. We found a progressive convergence of linguistic feature representation in native and nonnative listeners with increased proficiency, which enabled accurate decoding of language proficiency. At the same time, we found distinct neural signatures in native and nonnative listeners that persisted even in the most proficient listeners. This detailed view advances our understanding of the cortical processing of linguistic information in second-language learners and provides an objective measure of language proficiency.

Introduction

Instructed second-language (L2) learning is a challenging process that differs from native language (L1) acquisition. Learners who begin at a younger age generally achieve proficiency closer to that of a native speaker than learners who begin later in life[1,2]. In fact, adult learners rarely attain native-like L2 proficiency and instead carry-over L1 features to their L2 [3], which can have a major impact on their social lives[4–6]. Listening to L2 speech elicits stronger neural activations in shared linguistic areas[7,8] and engages cortical areas that are not active during L1 listening[9]. However, a consensus has yet to be reached on the precise neural underpinnings of L2 perception and how it differs from L1 processing (e.g., Perani and Abutalebi 2005; Bohn and Munro 2007; Marsden et al. 2013).

To elucidate the neural mechanisms that underlie L2 perception, it is crucial to assess the effect of proficiency on objective neural measures that capture the multifaceted cortical encoding of language. This is a complex task especially because speech perception involves the analysis of various acoustic and linguistic features, a process that is thought to engage a hierarchical neural network

composed of various interconnected cortical regions[12]. Distinct stages of processing were shown to be affected differently by proficiency, with some of them becoming more native-like than others for proficient L2 users. Part of the evidence comes from electro- and magneto-encephalography (EEG and MEG respectively) research, which showed the effect of proficiency at the levels of phonemes[13], syntax[14,15], and semantics[16]. These studies measured the changes in well-known event-related potentials, such as the MMN, N400, and P600. These approaches, however, use unnatural speech stimuli (e.g., isolated syllables or violative speech sentences) which do not fully and realistically activate the specialized speech cortex[17–19]. In addition, these approaches consider various levels of speech perception independently and in isolation. Language learning, on the other hand, involves simultaneous acquisition of novel phonetic contrasts[20,21], new syllabic structures (phonotactics)[22], and new words. A more complete view of the neural basis of language learning therefore requires a joint study of multiple levels of the linguistic hierarchy (**Fig. 1A**) to advance our understanding of L2 perception by informing us on the precise effect of proficiency on the cortical processing strategies that underpin sound and language perception[23–25].

Previous effort in using naturalistic speech stimuli to study language proficiency showed a modulation of EEG phase-synchronization during naturalistic speech listening both at sub-cortical (FFR[26,27]) and cortical levels (gamma EEG synchrony[28,29]). Specifically, stronger synchrony between EEG channels was shown for low-proficiency users[29], which is in line with theories positing that less experienced listeners require stronger cortical engagement, such as the *cortical efficiency theory*[29,30]. However, that work could not isolate neural signatures at particular linguistic stages. Recent studies have successfully isolated neural signatures of various linguistic levels based on speech-EEG synchrony (cortical tracking[31]) from a single electrophysiological recording. Such measures were derived based on low-frequency cortical responses to natural speech from audio-books[32–34] and cartoons[35,36], which were recorded non-invasively on both children and adults. Here, we adopted that framework to investigate how proficiency shapes the hierarchical cortical encoding in L2 subjects, and how that differs from L1 subjects. Our analysis focused on speech processing at the levels of sound acoustics[37,38], phonemes[32,34], phonotactics (statistics on phoneme

sequences[33,39]), and semantics[40–42]. We hypothesized that the neural encoding of all three levels of linguistic properties would be modulated by L2 proficiency, becoming more native-like without fully converging[43,44]. A different progression of this learning effect was expected for distinct linguistic levels. Specifically, we predicted phoneme and phonotactic responses, which benefit from but do not require sentence comprehension, to show a continuous progression starting from the earliest stages of learning, partly as a form of implicit learning[45]. Semantic-level encoding was also expected to show progressively stronger encoding[46], however with a most prominent change at an intermediate level of proficiency, when the comprehension of few words facilitates the understanding of the neighboring ones (e.g., semantic priming[47,48]), thus determining a turning point beyond which comprehension increases drastically.

To shed light on the neural mechanisms underlying the encoding of linguistic features, the present study combines objective neural indices of acoustic and linguistic processing to assess the differences between L1 and L2 subjects with varying proficiency levels during a natural speech listening task. Furthermore, we test the hypothesis that the hierarchical encoding of language presents fundamental differences between native and non-native listeners, which go beyond the changes due to proficiency alone.

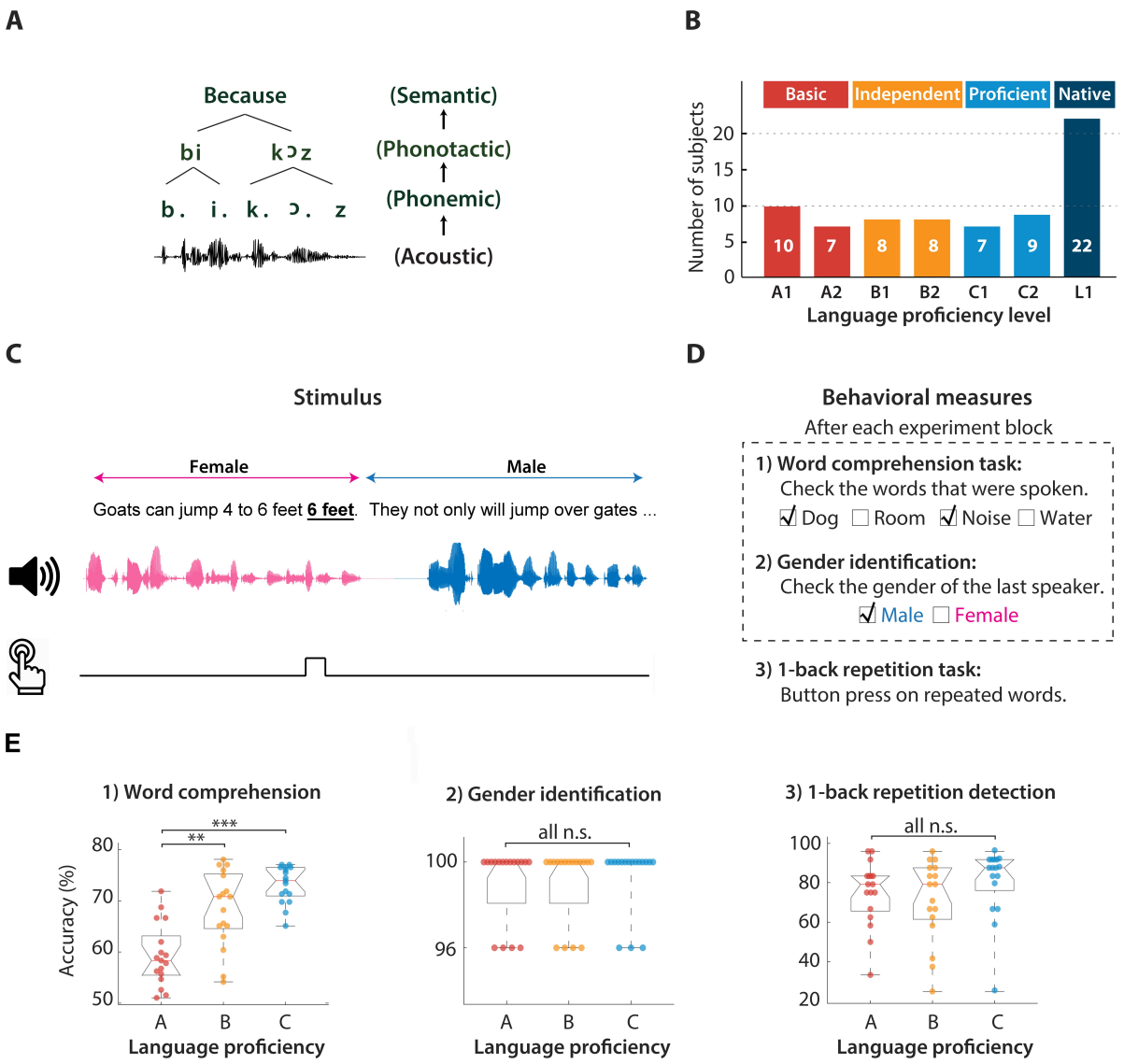


Figure 1. Experimental setup and behavioral results. *(A) Selected part of the speech processing hierarchy, from low-level acoustics (bottom) to higher-level linguistic properties (top). (B) Demographic distribution of the proficiency levels. (C)Sentences spoken by a male and a female speaker were presented in alternation. Participants were asked to detect a one-back phrase repetition (2-4 words), which occurred 1-5 times per experimental block, by pressing a button during the experimental block. (D) After each block, participants were asked to identify words that were spoken during the block from a list of eight, and to indicate the gender of the speaker at the end of the block. (E) Results for the three behavioral measures. Significant group differences (ANOVA, ** p<0.01, *** p < 0.001) were measured for the 'word comprehension' score that positively correlated with proficiency. No significant effects emerged for 'gender identification' and 'one-back tasks' as they are independent from the proficiency levels.*

Results

62-channel EEG was recorded from 73 participants as they listened to continuous speech sentences in English. 51 of them were native Standard Chinese (Mandarin) speakers with English as a non-native language. Two of these subjects were excluded because of synchronization issues with the EEG recordings. English proficiency was assessed by means of a standardized test of receptive skills that assigned participants to six different CEFR levels (Common European Framework of Reference for languages): *basic user* (A1 and A2 levels), *independent user* (B1 and B2 levels), and *proficient user* (C1 and C2 levels). The remaining 22 participants were native English speakers. The number of subjects in each group is shown in **Figure 1B**.

Subjects listened to 1.5 hours of speech material taken from a children's story book (Hank the Cowdog) narrated by four speakers (2 female). The gender of the speaker alternated for each consecutive sentence. The experiment was divided into twenty blocks, each about three minutes long, grouped in five sections (each composed of four blocks). To assess attention in L1 participants, three questions about the content of the story were asked after each block. All L1 participants were attentive and could answer at least 60% of the questions correctly. Because some L2 participants were not sufficiently proficient to understand the questions nor the speech, the subject's engagement and attention to the task were monitored with a modified, richer set of behavioral measures that could capture engagement even when the speech input was not understood (**Fig. 1C,D**). The first test was a word comprehension task, where at the end of each experimental block subjects were asked to select which words had been spoken in the last block from a list of eight. The second task, which did not require comprehension, was a gender identification task, where the subjects were asked to report the gender of the speaker that uttered the last sentence before the end of the experimental block. Finally, subjects did a one-back repetition detection task in which they were required to detect repeated phrases (containing 2-4 words and was always at the end of a sentence), which occurred 1-5 times per experimental block. Participants were instructed to use a clicker counter when they heard a repeated phrase and to report the number of clicks at the end of each

block. EEG data corresponding to the phrase repetition were excluded from the analysis to remove contamination due to the motor action. The detection of repeated sounds has been successfully used in many studies using non-speech and nonsense speech sounds [49–51]. The result of the three behavioral tasks is shown in **Figure 1E**. As expected, the word comprehension score increased significantly with proficiency (ANOVA, $F(1.8, 44.3) = 24.1$; $p = 5.4*10^{-8}$, post-hoc comparisons: $p_{A\text{-}vs\text{-}B} = 0.003$, $p_{B\text{-}vs\text{-}C} = 0.084$, $p_{A\text{-}vs\text{-}C} < 0.001$). However, all subjects were able to perform the gender identification and one-back tasks with similar accuracy across proficiency levels, suggesting a comparable degree of engagement among participants across groups (*gender identification task*: ANOVA, $F(1.4, 34.7) = 0.1$, $p = 0.90$; *one-back task*: ANOVA, $F(1.9, 45.9) = 1.1$; $p = 0.34$).

To investigate the low- *versus* higher-level brain processing of speech, we used linear regression models to measure the coupling between the low-frequency cortical signals (1–15 Hz) and progressively more abstract properties of the linguistic input. Specifically, we selected features describing acoustic properties based on the speech envelope (amplitude envelope – **Env**, spectrogram – **Sgr**, half-way rectified first derivative of the envelope **Env'** [52,53]), phoneme onsets (**Pon**), phonetic articulatory features (**Phn** [20,21]), its reduced version marking only vowels and consonants (**Pvc**), phonotactics (**Pt** [54]), and semantic dissimilarity (**Sem** [55,56]). Then, we used multiple linear regression to estimate the temporal response function (TRF) describing the spatio-temporal relationship between stimulus descriptors and EEG signal (**Figure 2** [57–59]). The combination of multiple speech features in a single multivariate model allowed us, for the first time, to assess the hierarchical processing of L2 speech from a single EEG recording session based on natural speech.

Hierarchical cortical encoding of native and non-native speech

Forward TRF models were fit between concatenations of speech features and the EEG signal for a broad time-latency window from 0 to 600ms to take into account various time dependencies at all hierarchical levels of interest. The stimulus descriptor included **Env**, **Env'**, **Pon**, **Pvc**, **Pt**, and **Sem** (**ALL;** see **Methods**). This combination of features allowed us to capture and discern EEG variance corresponding to various hierarchical stages, while using a low-dimensional descriptor (8

dimensions). We also fit TRF models with an extended stimulus descriptor (**EXT**) that includes **Sgr**, **Env'**, **Pon**, **Phn**, **Pt**, and **Sem**, which provided us with a higher level of detail in the spectrotemporal and phonological processing of speech. However, this increased dimensionality of the model (40 dimensions) makes fitting the model more challenging. Leave-one-out cross-validation indicated that the resulting TRF models could reliably predict the EEG signal for all subjects ($r_{ALL} > r_{ALL_SHUFFLE}$ and $r_{EXT} > r_{EXT_SHUFFLE}$, $p < 0.01$, permutation test where input sentences were randomly shuffled, N = 100; EEG prediction correlations were averaged across all electrodes).

Comparable cortical encoding of speech acoustics in L1 and L2 listeners

Figure 2B shows the model weights corresponding to the **Env** descriptor (part of TRF_{ALL}) after averaging across all electrodes and all subjects within each proficiency group (A, B, C, and L1). TRFs for the four groups appear temporally synchronized, which was expected for cortical responses to low-level acoustics. Effects of proficiency emerged on the TRF_{Env} magnitude that, however, did not survive correction for multiple comparison at any of the EEG channels (two-tailed permutation tests with FDR-correction over time-latencies, N = 100). In spite of that, the envelope response in L2 participants could have plausibly become more similar to that of native speakers with proficiency. To test for this possibility, we measured the Pearson's correlation scores between the TRF_{Env} for each L2 subject and the average TRF_{Env} weights across all L1 participants. This measure of similarity between L1 and L2 did not show any significant difference between A, B, and C groups ($p > 0.05$).

Although envelope TRFs have been proven robust and contributed to the study of various aspects of auditory perception[60–63], we have also modeled the low-level auditory responses by considering the acoustic spectrogram (**Sgr**), which was shown to be a better predictor of the EEG signal[32,64]. However, observing TRF_{Sgr} (part of TRF_{EXT}) for different auditory frequency bands did not lead to new clear-cut insights in this case, thus the rest of the manuscript will focus on the envelope TRF results.

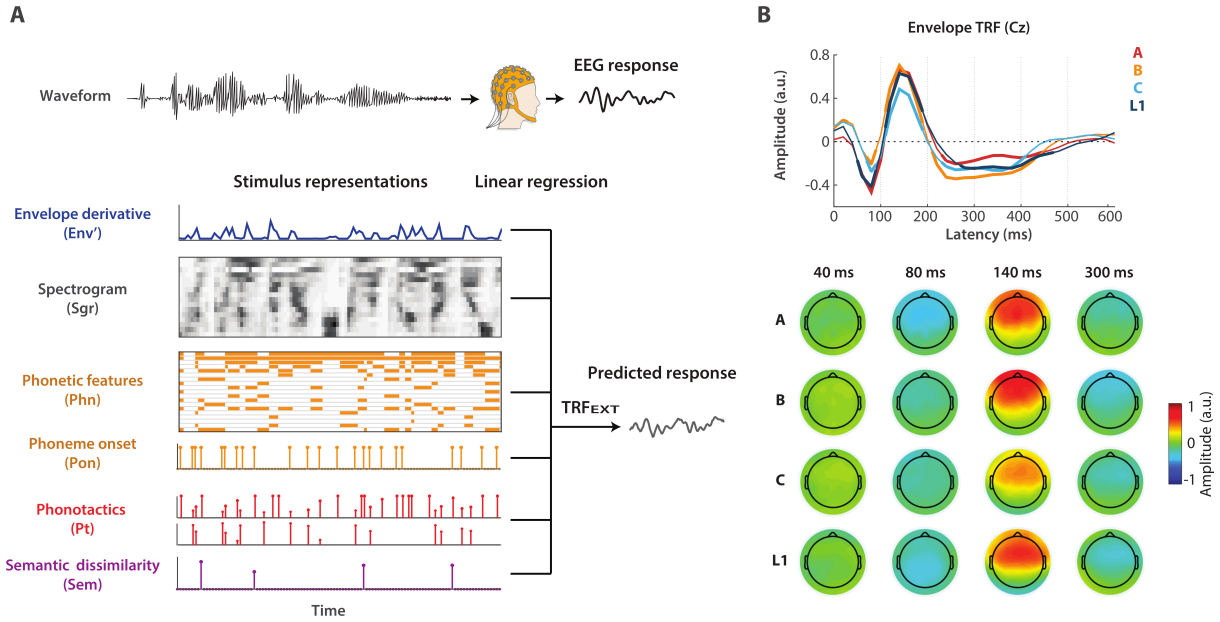


Figure 2. Investigating the hierarchical cortical encoding of language with the Temporal Response Function (TRF) analysis framework. *(A) Multichannel EEG signal was recorded as participants listened to audio-stories. Acoustic and linguistic information were extracted from the stimulus and encapsulated into data vectors and matrices. Multivariate linear regression was used to identify a linear fit that optimally predicted the EEG signal from features at multiple linguistic levels (EXT). The same procedure was also run on a more compact set of descriptors (ALL), which differed in that Sgr and Phn were replaced by Env (broadband envelope instead of a 16-band spectrogram) and Pvc (indicator variables for vowel and consonants only rather than for a 19-dimensional set of phonetic features) respectively. (B) Envelope TRF weights (TRF_{Env}) averaged across all EEG channels at peri-stimulus time-latencies from 0 to 600 ms. TRF_{Env} was part of a model that was fit by including features at all other levels of interest (ALL). Thick lines indicate weights that were statistically different from zero across all subjects of a group ($p < 0.05$, FDR corrected permutation test). The topographies indicate the distribution of TRF weights across channels for four selected time-latencies.*

Cortical encoding of phonemes in L2 listeners converges to L1 encoding with proficiency

Phonetic feature information was represented by the categorical descriptor **Phn**, which marked the occurrence of a phoneme with a rectangular pulse for each corresponding phonetic features (see **Methods**)[32]. TRFs were fit for each subject by combining the **Phn** descriptor with all others in the **EXT** feature-set. The weights corresponding to the descriptor of interest, TRF_{Phn}, were extracted from TRF_{EXT}. In this case, the other descriptors played the role of nuisance regressors, meaning that they reduced the impact of acoustic-, phonotactic- and semantic-level responses on TRF_{Phn}. To assess the effects of proficiency and nativeness, TRF weights were compared between L1 and L2 proficiency levels A, B, and C. First, TRF_{Phn} was rotated from feature to phoneme space by means of a linear transformation (see **Methods**; **Supplementary Figure 1**). Then, a classical multidimensional scaling (MDS) was used to project the TRF_{Phn} (phonemes were considered as

objects and time-latencies as dimensions) onto a 2D component space for each proficiency group (**Figures 3A**) where distances represent the discriminability of particular phonetic contrasts in the EEG signal. The result for each L2 proficiency group was then mapped to the average L1-MDS space by means of a Procrustes analysis. **Figure 3A** shows the average L1-L2 distance across all phonemes for each L2 participant, with blue and red fonts indicating phonemes for L1 and L2 participants respectively. Shorter L1-L2 distances were measured for increasing L2 proficiency levels (**Figure 3B**: ANOVA, $F(1.4, 54.1) = 22.8$; $p = 1.6*10^{-8}$), indicating a progressive L2-to-L1 convergence of the phoneme TRFs with proficiency.

To test whether this phoneme convergence was more pronounced for English phonemes that do not exist in Standard Chinese (the native language of L2 subjects), we measured the discriminability between selected pairs of phonemes based on TRF_{Phn} for each proficiency group. The phonemic discriminability was assessed by calculating the L2/L1 distance ratio of a given phonemic contrast. Using phonemes that can occur in minimal pairs (words differentiated by only one phoneme, e.g., “bat” /bæt/, “pat” /pæt/), and thus must be discriminable for proper comprehension, we selected ‘English-only contrasts’, which contain at least one phoneme that does not occur in Standard Chinese (e.g., TH). Such unknown phonemes have been shown to be perceived by L2 speakers as the closest existing phonemic neighbor in their L1, thus presenting challenges in discrimination[65,66]. We expected discriminability to increase with proficiency when considering phonemic contrasts that exist in English but not in Standard Chinese, thus reflecting the improved discrimination skills of L2 listeners. Our data is sensitive to this learning process, as we show in **Figures 3C** for six selected English-only contrasts (T vs. TH, D vs. DH, V vs. W, Z vs. DH, B vs. P, Z vs. S) which all show increased discriminability when comparing the A and C proficiency-level groups. As expected, this was not the case for phonemic contrasts belonging to both English and Standard Chinese languages (F vs. P, L vs. F, W vs. F, N vs. NG, M vs. N, L vs. NG) which did not show any consistent change with proficiency.

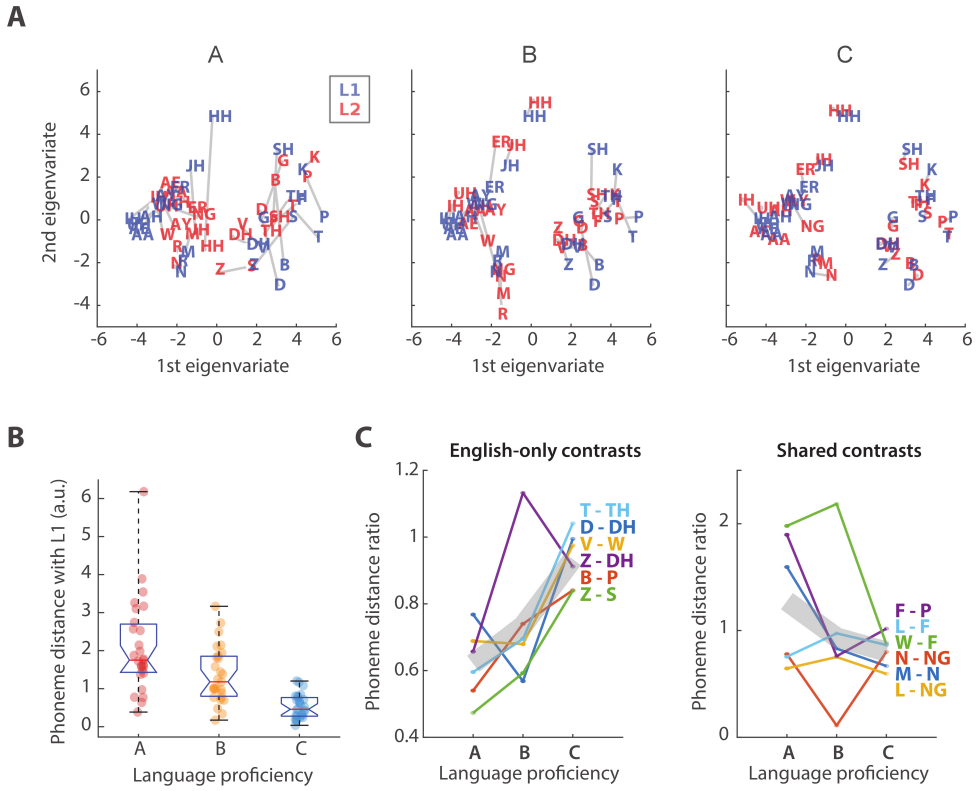


Figure 3. Effect of proficiency on L2 phoneme encoding. *(A) Phoneme distances for L1 and L2 are compared for each proficiency group by means of a multi-dimensional scaling (MDS) analysis based on the* TRF_{Ph} *weights at the electrode Cz and peri-stimulus time-latencies from 0 to 600 ms. Blue and red colors indicate phonemes for L1 and L2 participants respectively. (B) Distance between L1 and L2 phonemes for each language proficiency group. A significant effect of proficiency was measured on the L1-L2 phoneme distance (ANOVA,* $p = 1.6*10^{-8}$*). Error bars indicate the SE of the mean across phonemes. (C) Distance between phoneme-pairs for each proficiency level. The left panel shows results for contrasts existing in English but not in Standard Chinese, where we expected increasing discriminability with proficiency due to learning. The right panel shows distances for contrasts that exist both in English and Standard Chinese, where we did not expect a learning effect. Values were divided by the distance for L1 participants. Grey lines indicate the mean across all selected phonemic contrasts.*

Proficiency modulates phonotactic responses at both short and long latencies

In a given language, certain phoneme sequences are more likely to be valid speech tokens than others. The likelihood of a phoneme sequence $p_{1..n}$ being a valid speech token can be estimated with statistical models based on language-specific rules. One such model, called BLICK[54], returns the phonotactic *score* corresponding to the negative logarithm of the sequence likelihood. As such, that score has larger values for less likely (more surprising) sequences, providing us with a quantitative comparison between speech tokens. Crucially, this score is sensitive to changes in likelihood between the valid words composing natural conversational speech, thus allowing us to go beyond previous research based solely on phonotactic violations. Here, likelihood values for each phoneme

at position *k* of a word were calculated as the phonotactic score on each subsequence $p_{1..k}$, and used to modulate the amplitude of a phoneme onsets vector (all other time points were assigned to zero[39]; see **Methods**). **Figure 4A** compares the corresponding TRF weights (part of TRF_{ALL}) between proficiency groups at three scalp locations of interest. Qualitatively different TRF patterns emerged between groups, with an early positive component (~40 ms) that emerged consistently for all groups, an expected longer latency component (300-500 ms) that was less pronounced for L2 than L1 subjects, but it was significant for L2 with high and medium proficiency, and an unexpected earlier component (~120 ms) that emerged consistently only for all L2 groups but not L1 (FDR-corrected permutation tests). The same latencies showed significant effects of proficiency, which was measured as a Pearson's correlation between proficiency level and TRF magnitude. The topographical patterns in **Figure 4A** further clarify that this effect of proficiency was distributed across most scalp areas, but especially in centro-frontal scalp areas at 120 ms, while the effect at a latency of about 360 ms showed centro-parietal patterns.

Stronger and earlier cortical responses to semantic dissimilarity with proficiency

A similar analysis was conducted based on semantic dissimilarity rather than phonotactic scores. Specifically, content words were described according to a 400-dimensional feature space that was identified based on word co-occurrence (Word2Vec algorithm[55]). Then, semantic dissimilarity was quantified as the *distance* of a word with the preceding semantic context, thus resulting in a vector marking the onset of all content words with these distance values (see **Methods**)[40]. **Figure 4B** shows the semantic dissimilarity TRF_{Sem} (part of TRF_{ALL}) for three selected scalp channels. Average TRF_{Sem} for L1 participants were consistent with the results shown by Broderick and colleagues[40], with a negative component peaking at peri-stimulus latencies of 340-380 ms. Similar TRF patterns emerged for the L2 C-level participants, whose average TRF_{Sem} showed a negative component at comparable time-latencies, with trough at latencies between 340 and 440 ms (depending on the EEG channel). As expected, we observed significant correlations between TRF magnitude and proficiency over central and posterior scalp areas ($p < 0.05$, Pearson's correlation). Interestingly, an

unexpected significant bilateral centro-frontal negativity (BCN) peaking between 440 and 520 ms appeared in all L2 subjects but not in L1 subjects.

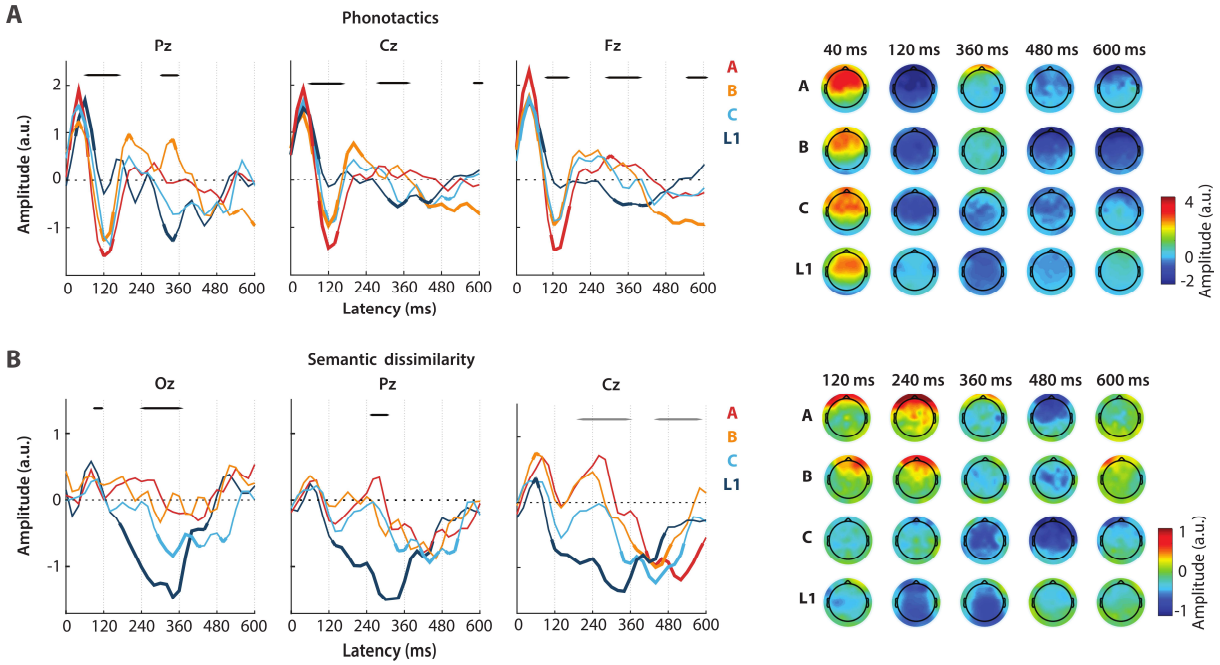


Figure 4. Effect of proficiency and nativeness on the EEG responses to phonotactics and semantic dissimilarity regressors. *(A) Model weights of the phonotactic TRF for three selected EEG channels (left) at peri-stimulus time-latencies from 0 to 600 ms. Results for distinct participant-groups are color-coded. Thick lines indicate weights that were statistically different from zero across all subjects of a group ($p < 0.05$, FDR corrected permutation test). Horizontal black lines indicate significant correlations between TRF weights and language proficiency ($p < 0.05$, Pearson's correlation). The topographies (right) indicate the TRF distribution across channels for five selected time-latencies. (B) Model weights of the semantic dissimilarity TRF for selected EEG channels (left) and selected time-latencies (right).*

Decoding language proficiency and nativeness

Our results indicate that language proficiency modulates the cortical responses at various linguistic processing levels. Given this relation, we examined the extent to which the proficiency of a subject could be predicted from the combined effects of different linguistic features. First, multilinear principal component analyses (MPCA) was conducted on the TRF weights corresponding to **Env**, **Phn**, **Pt**, and **Sem** separately, and the first component was retained for each of them. In doing so, information spacing along three dimensions (EEG channels, time latencies, and stimulus features – e.g., phonetic features) was compressed into a single value for each participant. A linear regression model was then fit to predict L2 proficiency (L1 subjects were excluded from this analysis) based on

those four TRF features. **Figure 5A** shows the effect of each regressor on the model fit (coefficient estimate and standard error), with an overall regression correlation $r = 0.68$. Significant effects were measured for each of the four features, and this was true also when the 'age' information and the 'one-back repetition detection' score (which was a measure of the attentional engagement to the experiment) were included in the regression fit. This result confirmed that the main effect of proficiency was not due to attention nor age.

A similar decoding approach was then used to assess whether and how robustly L2 proficiency could be decoded based on EEG indices of language processing. A set of 26 features was identified to most comprehensively describe the effects of L2 proficiency and nativeness on the TRFs. Features were based either on the TRF weights (as for **Figure 5A**), on the EEG prediction correlations based on subject-specific TRF models, or on EEG prediction correlations for each subject when using average TRF models that were fit on the other subjects, for A, B, C, and L1 groups separately (generic modelling approach[36,67]) (see **Methods** for a detailed list of features). Each of the 26 feature-vectors had multiple dimensions (e.g. electrodes, time-latency). For this reason, as it was described above, MPCA was used to reduce those vector to one-dimensional regressors. Support Vector Machine (SVM) regression was used to decode L2 proficiency based on such regressors. A backward elimination procedure identified a reduced set of features (**Supplementary Table 1**) whose combination produced optimal L2 proficiency decoding scores, with $MSE = 1.14$ and Pearson's correlation $r = 0.83$, $p = 3*10^{-13}$ (**Figure 5B**). Another way to quantify the quality of the proficiency decoding is to assess the A- vs. C-level classification by placing a simple threshold on the prediction values (value 3.5, which cuts the prediction space in half). This binary classification could identify A- vs. C-level participants with 91% accuracy.

Further analyses were conducted to assess the effect of nativeness on the EEG responses to speech. Specifically, differences in language processing between L1 and L2 subjects may be in part driven by a fundamental distinction between native and non-native language processing that is not due to proficiency *per se,* but rather due to differences in the L1 and L2 processing networks[68,69]. In

fact, the TRF results in **Figures 2-4** indicated that a higher proficiency level does not always lead to EEG responses that are equivalent to the ones of native speakers. Specifically, while there was some level of L1-L2 convergence for phoneme-level TRFs, this phenomenon was less pronounced for phonotactics and semantic dissimilarity responses, with marked differences between L1 and C-level L2 (e.g., the latency of the negative component at ~120 ms in TRF_{Pt}). Here, we attempted to disentangle those differences from the effect of L2 proficiency by conducting an SVM binary classification analysis for L1 versus L2 participants. This procedure used the same 26 features and backward elimination strategy as in the previous regression analysis. First, an L1 versus L2 classification accuracy of 87% was obtained when all 71 subjects were included in the analysis, with a baseline classification accuracy of 70% (95th percentile of a distribution of classification accuracy values when L1-L2 labels were randomly shuffled – 100 shuffles). Second, the same analysis was run on L1 and C-level L2 participants to assess more specifically the effect of nativeness. In this case, a classification accuracy of 73% was measured, with a baseline of 66%, thus indicating that the EEG responses to continuous speech reflect both the influence of L2 proficiency and nativeness.

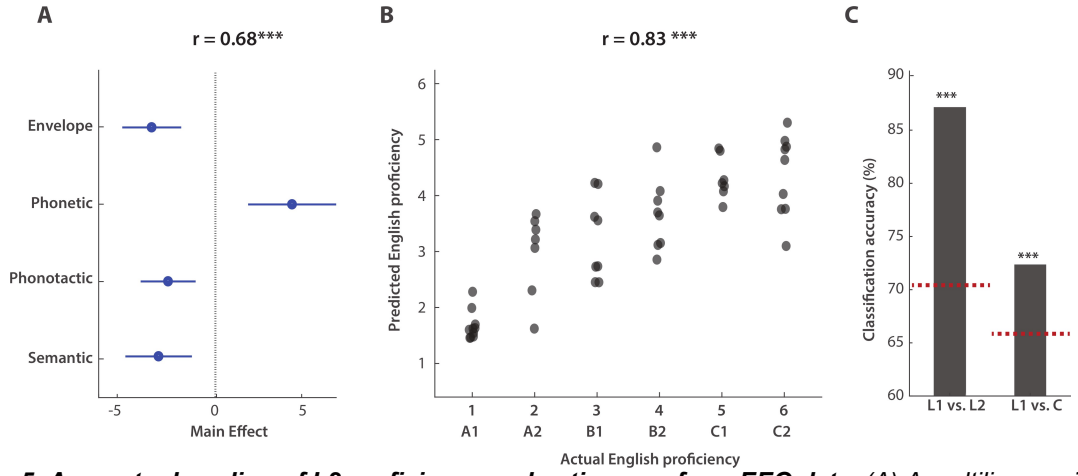


Figure 5. Accurate decoding of L2 proficiency and nativeness from EEG data. *(A) A multilinear principal component analysis (MPCA) was performed on the TRF weights corresponding to speech descriptors at all linguistic levels of interest. The first MPCA component was retained for the TRFs corresponding to Env, Phn, Pt, and Sem. The combination of those four features was predictive of L2 proficiency (r = 0.68), with significant effects for all features which were not due to group-differences in age or attention. (B) A support-vector regression analysis shows that EEG data accurately predict the L2 proficiency level at the individual subject level (r =0.83, MSE = 1.14). (C) Classification accuracy for L1 versus L2 and L1 versus C-level L2. The red dotted lines indicate the baseline classification levels, which were calculated as the 95th percentile of a distribution of classification accuracies derived after random shuffling the output class labels (N = 100).*

Discussion

The human brain responds differently when listening to second-language and native speech[13,15,16], typically leading to lower listening performances that vary between individuals and can be quantified with standardized language tests. Despite the general consensus for the cognitive, social, and economic advantages that come with high L2 proficiency, the neural underpinnings of second language perception and learning remain unclear[70,71]. One reason why this issue remains unresolved is methodological. Experimental evidence derived from direct neural measures is minimal and often limited to single linguistic properties[72–76], thus offering only a partial view of this complex brain mechanism. The present study establishes a methodological framework to provide a more comprehensive examination of the language processing system in naturalistic conditions. We isolated neural indices of speech perception at multiple processing stages from EEG responses to natural speech, revealing marked effects of L2 proficiency and nativeness that were robust at the individual subject level. Overall, the results confirm our hypothesis that the cortical encoding of spoken language in L2 listeners becomes progressively more similar to L1 with proficiency, and that this convergence is only partial as there are differences between L1 and L2 that persist even for listeners with high L2 proficiency. This finding suggests that proficiency and nativeness have somewhat distinct effects on the EEG signal, pointing to precise spatio-temporal neural patterns that may underpin the two phenomena.

Previous studies that investigated L2 perception in naturalistic paradigms focused on the relationship between neural activity and acoustic envelope, and found stronger coupling in L2 than L1 subjects[77]. That EEG result, which was found with a selective attention listening task in a multi-talker scenario, pointed to a link between increased listening effort and stronger cortical tracking of the speech envelope. However, it remained unclear which of the linguistic and non-linguistic properties of speech correlate with the acoustic envelope results in this increased cortical tracking. In fact, an increased coupling between EEG and speech envelope could reflect increased encoding of acoustic features, stronger reliance on higher order processes, or even activation of distinct cortical areas. For example, recent work indicated that envelope tracking increases with age due to a stronger

engagement of higher order areas, thus reflecting a difference in processing strategy for older listeners[78,79]. While our results did not show differences in envelope tracking (EEG prediction correlations) with proficiency, the shape of the envelope TRF significantly contributed to both L2 proficiency and nativeness decoding (**Figure 5**), which is in line with a link between acoustic-level encoding and effort. Interestingly, this result was obtained on a single-talker task with no competing noise. Therefore, the application of this same approach on a more cognitively demanding task[62,80] could help tease apart the effects of L2 proficiency and listening effort on the cortical encoding of acoustic features.

As we had hypothesized, the cortical encoding of phonemes changed with L2 proficiency, becoming progressively more similar to L1, which is in line with perceptual theories such as the expanded Native Language Magnet Theory (NML-e[81]) and the Perceptual Assimilation Model (PAM-L2[82]). Our TRF analysis has discerned individual phonemic contrasts, showing that the cortical encoding of phonemes becomes progressively more sensitive to contrasts existing in English but not Standard Chinese (**Figure 3**). This work extends previous findings on the cortical encoding of phonemes[32,34,64] by demonstrating that EEG responses to natural speech show sensitivity to individual phoneme contrasts, with response patterns that become progressively more categorical with proficiency. Furthermore, that result goes beyond previous work[64] by revealing a low-frequency EEG component that could not be explained by simple acoustic features such as acoustic envelope, derivative of the envelope, and spectrogram. Our results are in line with the majority of theories on L2 perception which suggest the impact of a subject's L1 on phonological encoding of the L2. Specifically, **Figure 3A** indicates that the native language constitutes a "starting point" for the phonological encoding in L2 learners, which then changes with experience and converges toward the encoding for L1 listeners. Reproducing this work on participants with other native languages could provide us with detailed insights on the effect of the native language on the phoneme encoding in high proficiency L2 learners. Furthermore, additional data based on a more balanced design, with subjects listening to both their native and non-native language, could reveal whether and how learning a particular L2

influences the cortical processing of the native language[19,83], as was postulated by the bidirectional cross-linguistic influence principle in the Speech Learning Model (SLM[84]).

Proficiency was also shown to shape language encoding at the level of phonotactics, with TRFs in L2 subjects progressively converging toward L1 TRFs. Our results indicate two effects of phonotactics. First, an hypothesized, TRF component peaking at speech-EEG latencies of about 300-450ms, as measured in a previous EEG study by our group[39], with more negative responses for higher proficiency-levels (**Figure 4A**); Second, an effect at shorter latencies of about 120ms, where a negative component that was not present for L1 participants emerged for L2 participants. Interestingly, a component reflecting phonotactics was previously measured at that speech-neural signal latencies with MEG[33] but not EEG. Our finding provides a new link between the EEG and MEG literature by clarifying that phonotactics modulates EEG responses at both shorter and longer latencies, and that the effect at shorter latencies emerges for L2 learners but not L1. This discrepancy may be due to the difference in the type of signal recorded by EEG and MEG modalities. The larger values for lower-proficiency users could reflect an effect of surprise on the phoneme sequences due to the use of an incorrect (or imperfect) model of phonotactics.

Semantic dissimilarity TRFs were previously shown to be characterized by a negative centro-parietal component at speech-EEG latencies of about 350-400ms. This finding is in line with previous work on the N400 event-related response[42,85,86] which shows that this component is modulated by intelligibility and attention[40]. Similarly, we expected a strong response negativity for users with higher language proficiency, and no response for people with no English at all. Consistent with this hypothesis, our results identified a posterior component with magnitude that increases with proficiency (**Figure 4B**). In addition, an unexpected centro-frontal component arose at latencies of about 440-520ms which was negatively correlated with the latency of response rather than the magnitude of the component. This bilateral centro-frontal negativity (BCN) emerged even for participants with no English understanding, thus reflecting neural correlates time-locked to word onset but not semantics *per se*. This component may instead be related to other processes, such as the processing of sentence structure, memory, and learning of frequent words[87–90]. Further work is

needed to clarify whether that signal reflects, for example, the familiarity with particular words, or it is related to ERP components such as the left anterior negativity (LAN), which was shown to reflect processing difficulties in morpho-syntax[91,92].

Although both phonotactics and semantic level TRFs for L2 showed some level of convergence to L1, there was also a pronounced difference between L1 and C-level L2 participants, which was also reflected in the significant L1 versus C classification result in **Figure 5C.** This effect may reflect fundamental differences in the cortical mechanisms underlying L1 and L2 processing, rather than an effect of proficiency *per se*. This effect of nativeness that is somewhat different from the effect of proficiency is in line with the observation that a second language learnt after a certain critical (or sensitive) period usually leads to lower language proficiency than those of a native speaker[1,93,94]. More data could provide further insights on this topic, for example, by comparing L1 monolinguals with bilinguals and multilinguals with a wide range of learning-onsets of the English language. Further research is also needed to better understand the precise sources of the effect of nativeness, e.g. by comparing L1 and high-proficiency L2 listeners with a semantic task guaranteeing the same level of comprehension for all participants. Such a task could not be employed in the present study, whose primary focus was the effect of proficiency across L2 participants from A- to C-levels which, by design, presented variable levels of comprehension.

Our analysis focused on just few components of the speech processing hierarchy, namely the acoustic, phonemic, phonotactic, and semantic levels. One powerful element of this framework is that it can be extended to other levels of processing without the need for additional data. In fact, the EEG responses to natural speech likely reflect many more components of interest than the ones targeted in this occasion, and demonstrating how to isolate them would give us insights on each newly added feature and its link with proficiency, as well as giving us the chance to improve the accuracy of our EEG-based L2 proficiency assessment. For these reasons, we believe that a wide collaborative effort under a common protocol of data acquisition with EEG/MEG and natural stimuli could significantly and quickly advance our understanding of the speech and language cortical processing network (and could indeed extend to other questions of interest).

Understanding the neural underpinnings of second language perception and learning becomes particularly relevant when we consider that there are more children throughout the world that have been educated via a second (or a later acquired) language rather than exclusively via their L1[95]. Furthermore, there is evidence for perceptual advantage of bilinguals and multilinguals that is due to cross-language transfer[83,96,97], and particular combination of languages may be better than others in the emergence of such a benefit. Further work in this direction may provide us with tools to predict the perceptual advantage that a particular second language would bring to a person given their background, thus constituting the basis for a procedure that, for example, could inform us on what second languages should be encouraged in school to particular individuals.

Materials and Methods

Participants

Fifty-one healthy subjects (twenty-four male, aged between 18 and 60, with median = 24 and mean = 27.5, forty-eight were right-handed) that learnt English as a second language (or that did not speak English) participated in the EEG experiment (L2 group). Two of these subjects were excluded because of issues with the EEG recordings (data could not be synchronized because of missing trigger signals). L2 participants were asked to take a 20-minute American English listening level test before the experiment. According to the results of this assessment, each participant was assigned to one of seven proficiency groups according to the CEFR framework (Common European Framework of Reference for Languages): A1, A2, B1, B2, C1, C2 (from low to high proficiency). A-, B-, and C- levels indicated basic, independent, and proficient users respectively. The A1 group included participants with no English understanding.

Data from twenty-two native English speakers (twelve male, aged between 18 and 45, twenty were right-handed) was originally collected for a previous study[34] based on the same experiment, with the same setup and location (L1 group). All subjects reported having normal hearing and no history of

neurological disorders. All subjects provided written informed consent and were paid for their participation. The Institutional Review Board of Columbia University at Morningside Campus approved all procedures.

Stimuli and experimental procedure

EEG data were collected in a sound-proof, electrically shielded booth in dim light conditions. Participants listened to short stories narrated by two speakers (1 male) while minimizing motor movements and maintaining visual fixation on a crosshair at the center of the screen. The male and the female narrators were alternated to minimize speaker-specific electrical effects. Stimuli were presented at a sampling rate of 44,100 Hz, monophonically, and at a comfortable volume from loudspeakers in front of the participant. Each session consisted of 20 experimental blocks (3 min each), divided in five sections that were interleaved by short breaks. Participants were asked to attend to speech material from seven audio-stories that were presented in a random order. To assess attention in L1 participants, three questions about the content of the story were asked after each block. All L1 participants were attentive and could all answer correctly at least 60% of the questions. A modified version of the behavioral task was used for L2 participants to assess their attention level, as some L2 participants were not sufficiently proficient to understand the questions nor the speech sentences. Participants were asked three questions at the end of each block. First, we asked whether the last sentence of the section was spoken by a male or female speaker. Next, participants were asked to identify 3-5 words with high-frequency in the sentence from a list of eight words. Third, participants performed a phrase-repetition detection task. Specifically, the last two to four words were repeated immediately after the end of some of the sentences (1-5 per block). Given that our target was monitoring attention, a finger-tip clicker was used to count the repetitions so that they would be engaged in a detection and not counting task, which would instead require additional memory resources and, potentially, reduce their engagement to the main listening task. Participants were asked to indicate how many sentences in the story presented these repetitions at the end of each block.

EEG recordings and preprocessing

EEG recordings were performed using a g.HIamp biosignal amplifier (Guger Technologies) with 62 active electrodes mounted on an elastic cap (10 –20 enhanced montage). EEG signals were recorded at a sampling rate of 2 kHz. An external frontal electrode (AFz) was used as ground and the average of two earlobe electrodes were used as reference. EEG data were filtered online using a high-pass Butterworth filter with a 0.01 Hz cut-off frequency to remove DC drift. Channel impedances were kept below 20 kΩ throughout the recording.

Neural data were analyzed offline using MATLAB software (The Mathworks Inc). EEG signals were digitally filtered between 1 and 15 Hz using a Butterworth zero-phase filter (order 2+2 and implemented with the function *filtfilt*), and down-sampled to 50 Hz. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation.

Speech features

In the present study, we have assessed the coupling between the EEG data and various properties of the speech stimuli. These properties were extracted from the stimulus data based on previous research. First, we defined a set of descriptors summarizing *low-level acoustic properties* of the music stimuli. Specifically, a time-frequency representation of the speech sounds was calculated using a model of the peripheral auditory system[98] consisting of three stages: (1) a cochlear filter-bank with 128 asymmetric filters equally spaced on a logarithmic axis, (2) a hair cell stage consisting of a low-pass filter and a nonlinear compression function, and (3) a lateral inhibitory network consisting of a first-order derivative along the spectral axis. Finally, the envelope was estimated for each frequency band, resulting in a two dimensional representation simulating the pattern of activity on the auditory nerve[99] (the relevant Matlab code is available at *https://isr.umd.edu/Labs/NSL/Software.htm*). This *acoustic spectrogram* (**S**) was then resampled to 16 bands[32,100]. A *broadband envelope* descriptor (**E**) was also obtained by averaging all envelopes

across the frequency dimension. Finally, the *half-way rectified first derivative of the broadband envelope* (**E'**) was used as an additional descriptor, which was shown to contribute to the speech-EEG mapping and was used here to regress out the most acoustic-related responses as possible[64].

Additional speech descriptors were defined to capture neural signatures of higher-order speech processing. The speech material was segmented into time-aligned sequences of phonemes using the Penn Phonetics Lab Forced Aligner Toolkit[101], and the phoneme alignments were then manually corrected using Praat software[102]. *Phoneme onset* times were then encoded in an appropriate univariate descriptor (**Pon**), where ones indicate an onset and all other time samples were marked with zeros. An additional descriptor was also defined to distinguish between *vowels and consonants* (**Pvc**). Specifically, this regressor consisted of two vectors, similar to **Pon**, but marking either vowels or consonants only. While this information was shown to be particularly relevant when describing the cortical responses to speech[33], there remains additional information on phoneme categories that contributes to those signals[32,103]. This information was encoded in a 19-dimensional descriptor indicating the *phonetic articulatory features* corresponding to each phoneme (**Phn**). Features indicated whether a phoneme was voiced, unvoiced, sonorant, syllabic, consonantal, approximant, plosive, strident, labial, coronal, anterior, dorsal, nasal, fricative, obstruent, front (vowel), back, high, low. The **Phn** descriptor encoded this categorical information as step functions, with steps corresponding to the starting and ending time points for each phoneme.

Next, we encoded *phonotactic probability* information in an appropriate two-dimensional vector (**Pt**)[33,39]. Probabilities were derived by means of the BLICK computational model[54], which estimates the probability of a phoneme sequence to belong to the English language. This model is based on a combination of explicit theoretical rules from traditional phonology and a maxent grammar[104], which find optimal weights for such constraints to best match the phonotactic intuition of native speakers. The phonotactic probability was derived for all phoneme sub-sequences within a word ($ph_{1..k}$, $1 \leqslant k \leqslant n$, where n is the word length) and used to modulate the magnitude of a phoneme onset vector ($\mathbf{Pt_1}$). A second vector was produced to encode the change in phonotactic probability due to the addition of a phoneme ($ph_{1..k}$ - $ph_{1..k-1}$, $2 \leqslant k \leqslant n$) ($\mathbf{Pt_2}$).

Finally, a semantic dissimilarity descriptor was calculated for content words using word2vec[55,105], a state-of-the-art algorithm consisting of a neural network for the prediction of a word given the surrounding context. In this specific application, a sliding-window of 11 words was used, where the central word was the output and the surrounding 10 words were the input. This approach is based on the "distributional hypothesis" that words with similar meaning occur in similar contexts, and it uses an artificial neural network approach to capture this phenomenon. This network has a 400-dimensional hidden layer that is fully connected to both input and output. For our purposes, the weights of this layer are the features used to describe each word in a 400-dimensional space capturing the co-occurrence of a content word with all others. In this space, words that share similar meanings will have a closer proximity. The semantic dissimilarity indices are calculated by subtracting from 1 the Pearson's correlation between a word's feature vector and the average feature vector across all previous words in that particular sentence (the first word in a sentence was instead correlated with the average feature vector for all words in the previous sentence). Thus, if a word is not likely to co-occur with the other words in the sentence, it should not correlate with the context, resulting in a higher semantic dissimilarity value. The semantic dissimilarity vector (**Sem**) marks the onset of content words with their semantic dissimilarity index.

Computational model and data analysis

A single input event at time t_0 affects the neural signals for a certain time-window [t_1, t_1+t_{win}], with $t_1 \geq 0$ and $t_{win} > 0$. Temporal response functions (TRF) were fit to describe the speech-EEG mapping within that latency-window for each EEG channel (TRF[106,107]). We did this by means of a regularized linear regression[58] that estimates a filter that allows to optimally predict the neural response from the stimulus features (forward model; **Fig. 2A**). The input of the regression also included time-shifted versions of the stimulus features, so that the various time-lags in the latency-window of interest were all simultaneously considered. Therefore, the regression weights reflect the relative importance between time-latencies to the stimulus-EEG mapping and were here studied to infer the temporal dynamics of the speech responses (see **Figures 2** and **3**). Here, a time-lag window of 0–600 ms

was used to fit the TRF models which is thought to contain most of the EEG responses to speech of interest. The reliability of the TRF models was assessed using a leave-one-out cross-validation procedure (across trials), which quantified the EEG prediction correlation (Pearson's *r*) on unseen data while controlling for overfitting. Note that the correlation values are calculated with noisy EEG signal, therefore the *r*-scores can be highly significant even though they have low absolute values ($r \sim 0.1$ for sensor-space low-frequency EEG[32,64,100]).

Stimulus descriptors at the levels of acoustics, phonemes, phonotactics, and semantics were combined in a single TRF model fit procedure. This strategy was adopted with the goal of discerning EEG responses at different processing stages. In fact, larger weights are assigned to regressors that are most relevant for predicting the EEG. For example, a TRF derived with **Pt** alone could reflect EEG responses to phonotactics and phoneme onset. A TRF based on the combination of **Pt** and **Pon** would instead discern their respective EEG contributions, namely by assigning larger weights to **Pt** for latencies that are most relevant to phonotactics. Here, individual-subject TRFs were fit by combining **Env**, **Env'**, **Pvc**, **Pon, Pt**, and **Sem** (stimulus descriptor **ALL**). We also fit TRF models with an extended stimulus descriptor (**EXT**) including **Sgr**, **Env'**, **Phn**, **Pon, Pt**, and **Sem**, which provided us with a higher level of detail on spectrotemporal and phonological speech features, at the cost of higher dimensionality. The combined stimulus descriptor had 40 dimensions, which have to be multiplied by the number of time-lags (30 with sampling frequency 50Hz) to have the dimensionality of the TRF input. For this reason, we conducted all analysis on the reduced stimulus set **ALL**, while the **EXT** descriptor was used to assess spectrotemporal and phoneme TRFs.

The TRF weights constitute good features to study the spatio-temporal relationship between a stimulus feature and the neural signal. However, studying this relationship for multivariate speech descriptor, such as **Phn**, requires the identification of criteria to combine multiple dimensions of TRF weights. One solution to use the EEG prediction correlation values to quantify the goodness of fit for a multivariate TRF model. Here, we considered the relative enhancement in EEG prediction correlation when **Phn** was included in the model (using the **ALL** feature-set), thus allowing us to discern the relative contribution of phonetic features to the neural signal. This isolated index of

phoneme-level processing was also shown to correlate with psychometric measures of phonological skills[35]. Additional analyses were conducted with a generic modelling approach[108]. Specifically, one generic TRF model was derived for each of the groups A, B, C, and L1 by averaging the regression weights from all subjects within the group. Then, EEG data from each left-out subject (whose data was not included in the generic models) was predicted with the four models. The four prediction correlations were used as indicators of how similar the EEG signal from a subject was to the one expected for each of the four groups, providing us with a simple classifier of proficiency and nativeness.

Proficiency-level decoding

Support Vector Regression (SVR) with radial basis function kernel was used to decode the proficiency level of L2 participants. The output of the regression is the proficiency level, a continuous integer variable with six possible values corresponding with A1, A2, B1, B2, C1, and C2. The input of the SVR was the concatenation of 26 features derived from the TRF analysis described in the previous section. Specifically, we chose features based on the TRF weights (9 features), subject-specific EEG prediction correlations (5 features), and generic models EEG prediction correlations (12 features).

Each feature had multiple dimensions, such as EEG electrodes and time-latencies. A multilinear principal component analysis (MPCA) was performed to summarize each of them with a single value. The first component was retained for the TRFs corresponding to envelope, phoneme onsets, phonetic feature, phonotactics, and semantic dissimilarity. Based on previous TRF studies and our initial hypotheses, we complemented the result of this lossy compression by adding distinctive feature that summarize specific aspects of interest of the TRFs. For speech acoustics, we included information on the power spectrum of the TRF (the EEG responsiveness to 16 logarithmically-spaced sound frequencies) by collapsing the weights in TRF_{ALL} corresponding to **Sgr** across the time-latency dimension. MPCA was then run on the resulting values to quantify this spectral feature with a single value per subject. For phonotactics and semantic dissimilarity, the strength of the main

TRF components was summarized by averaging the regression weights over selected time-windows and electrodes where they were strongest (80-140 and 300-700 for **Pt**, and 300-700 for **Sem**, at Fz, Cz, and Oz respectively).

EEG prediction correlations were calculated when TRF models were trained and tested on each participant separately (with leave-one-out cross-validation across recording blocks). This procedure provided us with a correlation score for each electrode, which was then summarized with a single value by running MPCA and retaining the first component. This procedure provided us with four features for EEG predictions based on **Env**, **Phn**, **Pt**, **Sem**. A fifth feature was derived by measuring the increase in EEG prediction correlations when **Phn** was included in the stimulus set by concatenating it with **Env'** and **Sgr** (**Phn_r**). This subtraction was suggested to constitute an isolated measure of phoneme-level processing[32,35]. Finally, EEG signals from a subject were also predicted with TRF models fit on all other subjects, grouped in **A**, **B**, **C**, and **L1**, with the rationale that the EEG data from a given subject should be best predicted by TRF models from subjects of the same group. This approach, which has been referred to as average models or generic modelling approach[36,108], provided each subject with a score for each group and for each feature of interest. Here we selected **Env'**, **Phn_r**, and the concatenation of **Pt** and **Sem**. MPCA was then used for dimensionality reduction as for the other features, providing us with twelve features (4 groups and 3 predicting features).

SVR was used to decode the L2 proficiency level, for the binary classification L1 versus L2, or for the binary classification L1 versus C-level L2 with leave-one-out cross-validation. A backward elimination procedure was used to identify that optimal set of features that minimize the mean squared error (MSE) of the decoded proficiency levels. Specifically, starting from a set containing all the features, the regressor whose exclusion produced the larger decrease in MSE was removed at each step. This procedure continued as long as there was at least 5% improvement on the MSE score (please see **Supplementary Table 1** for a full list of features and information on the selected feature for the L2 decoding and L1 vs. L2 classification procedures).

Statistical analysis

Statistical analyses were performed using two-tailed permutation tests for pair-wise comparisons. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach. One-way ANOVA was used to assess when testing the significance of an effect over multiple (> 2) groups. The values reported use the convention $F(df, df_{error})$. Greenhouse-Geisser corrections was applied when the assumption of sphericity was not met (as indicated by a significant Mauchly's test). FDR-corrected Wilcoxon tests were used after the ANOVA for post-hoc comparisons.

Author Contributions

The study was conceived by J.N., N.M., and G.D.L.; the experiments were designed by J.N. and B.K. programmed the tasks; J.N. collected the EEG data; G.D.L., N.M., J.N., J.Y. analyzed the data; G.D.L., J.N., N.M. wrote the first draft of the manuscript; J.Y., B.K., and S.S. edited the manuscript.

References

1. Hartshorne, J. K., Tenenbaum, J. B. & Pinker, S. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* **177,** 263–277 (2018).

2. Kuhl, P. K. Early Language Learning and Literacy: Neuroscience Implications for Education. *Mind, Brain, Educ.* **5,** 128–142 (2011).

3. Kotz, S. A. A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain Lang.* **109,** 68–74 (2009).

4. Lev-Ari, S. & Keysar, B. Why don't we believe non-native speakers? The influence of accent on credibility. (2010). doi:10.1016/j.jesp.2010.05.025

5. DeJesus, J. M., Hwang, H. G., Dautel, J. B. & Kinzler, K. D. Bilingual children's social preferences hinge on accent. *J. Exp. Child Psychol.* **164,** 178–191 (2017).

6. Mayo, L. H., Florentine, M. & Buus, S. Age of second-language acquisition and perception of speech in noise. *J. Speech, Lang. Hear. Res.* **40,** 686–693 (1997).

7. Yokoyama, S. *et al.* Cortical activation in the processing of passive sentences in L1 and L2: An fMRI study. *Neuroimage* **30,** 570–579 (2006).

8. Perani, D. & Abutalebi, J. The neural basis of first and second language processing. *Current Opinion in Neurobiology* **15,** 202–206 (2005).

9. Kim, K. H. S., Relkin, N. R., Lee, K. M. & Hirsch, J. Distinct cortical areas associated with native and second languages. *Nature* **388,** 171–174 (1997).

10. Marsden, E., Mitchell, R. & Myles, F. *Second language learning theories*. (2013).

11. Bohn, O.-S. & Munro, M. J. *Language experience in second language speech learning : in honor of James Emil Flege*. *Language learning and language teaching,* (2007).

12. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat Rev Neurosci* **8,** 393–402 (2007).

13. White, E. J., Genesee, F. & Steinhauer, K. Brain Responses before and after Intensive Second Language Learning: Proficiency Based Changes and First Language Background Effects in Adult Learners. *PLoS One* **7,** (2012).

14. Weber-Fox, C. M. & Neville, H. J. Maturational Constraints on Functional Specializations for Language Processing: ERP and Behavioral Evidence in Bilingual Speakers. *J. Cogn. Neurosci.* **8,** 231–56 (1996).

15. Hanna, J., Shtyrov, Y., Williams, J. & Pulvermüller, F. Early neurophysiological indices of second language morphosyntax learning. *Neuropsychologia* **82,** 18–30 (2016).

16. Ojima, S., Nakata, H. & Kakigi, R. An ERP study of second language learning after childhood: effects of proficiency. *J. Cogn. Neurosci.* **17,** 1212–28 (2005).

17. Van Petten, C. & Kutas, M. Interactions between sentence context and word frequencyinevent-related brainpotentials. *Mem. Cognit.* **18,** 380–393 (1990).

18. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* **18,** 903–911 (2015).

19. Zinszer, B. D., Chen, P., Wu, H., Shu, H. & Li, P. Second language experience modulates neural specialization for first language lexical tones. *J. Neurolinguistics* **33,** 50–66 (2015).

20. Chomsky, N. & Halle, M. The sound pattern of English. (1968).

21. Ladefoged, P. & Johnson, K. A course in phonetics. (2014).

22. Trubetzkoy, N. S. Principles of Phonology. (1969).

23. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31,** 32–59 (2016).

24. Huettig, F. & Mani, N. Language, Cognition and Neuroscience Is prediction necessary to understand language? Probably not. (2015). doi:10.1080/23273798.2015.1072223

25. Federmeier, K. D. Thinking ahead: The role and roots of prediction in language

comprehension. *Psychophysiology* **44,** 491–505 (2007).

26. Krizman, J., Slater, J., Skoe, E., Marian, V. & Kraus, N. Neural processing of speech in children is influenced by extent of bilingual experience. *Neurosci. Lett.* **585,** 48–53 (2015).

27. Omote, A., Jasmin, K. & Tierney, A. Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex* **93,** 146–154 (2017).

28. Reiterer, S., Pereda, E. & Bhattacharya, J. Measuring second language proficiency with EEG synchronization: how functional cortical networks and hemispheric involvement differ as a function of proficiency level in second language speakers. *Second Lang. Res.* **25,** 77–106 (2009).

29. Reiterer, S., Pereda, E. & Bhattacharya, J. On a possible relationship between linguistic expertise and EEG gamma band phase synchrony. *Front. Psychol.* **2,** (2011).

30. Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M. & Tohkura, Y. Effects of language experience: Neural commitment to language-specific auditory patterns. *Neuroimage* **26,** 703–720 (2005).

31. Obleser, J. & Kayser, C. Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences* **23,** 913–926 (2019).

32. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* **25,** (2015).

33. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Curr. Biol.* **28,** 3976–3983.e5 (2018).

34. Khalighinejad, B., Cruzatto da Silva, G. & Mesgarani, N. Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *J. Neurosci.* (2017). doi:10.1523/jneurosci.2383-16.2017

35. Di Liberto, G. M. *et al.* Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. *Neuroimage* **NIMG-17-29,** 70–79 (2018).

36. Jessen, S., Fiedler, L., Münte, T. F. & Obleser, J. Quantifying the individual auditory and visual

brain response in 7-month-old infants watching a brief cartoon movie. *Neuroimage* **202,** 116060 (2019).

37. Lalor, E. C. & Foxe, J. J. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.* **31,** 189–193 (2010).

38. Ahissar, E. *et al.* Speech Comprehension is Correlated with Temporal Response Patterns Recorded from Auditory Cortex. *Proc Natl Acad Sci U S A* **98,** 13367–13372 (2001).

39. Di Liberto, G. M., Wong, D., Melnik, G. A. & de Cheveigne, A. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage* **196,** 237–247 (2019).

40. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Curr. Biol.* (2018). doi:10.1016/j.cub.2018.01.080

41. Hagoort, P. & Brown, C. M. ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia* **38,** 1518–1530 (2000).

42. Kutas, M. & Federmeier, K. D. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62,** 621–47 (2011).

43. Selinker, L. Interlanguage. *IRAL - Int. Rev. Appl. Linguist. Lang. Teach.* **10,** 209–232 (1972).

44. Han, Z. H. *Fossilization in adult second language acquisition*. *Fossilization in Adult Second Language Acquisition* (2004). doi:10.25264/2519-2558-2019-6(74)-150-153

45. Conway, C. M., Bauernschmidt, A., Huang, S. S. & Pisoni, D. B. Implicit statistical learning in language processing: Word predictability is the key. *Cognition* **114,** 356–371 (2010).

46. McLaughlin, J., Osterhout, L. & Kim, A. Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nat. Neurosci.* **7,** 703–704 (2004).

47. Osterhout, L. & Holcomb, P. Event - Related Potentials and Language. in *Electrophysiology of the Mind: Event - Related Brain Potentials and Cognition* 171–187 (1995).

48. Kellenbach, M. L., Wijers, A. A. & Mulder, G. Visual semantic features are activated during the processing of concrete words: Event-related potential evidence for perceptual semantic priming. *Cogn. Brain Res.* **10,** 67–75 (2000).

49. Evans, S. & Davis, M. H. Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cereb. Cortex* **25,** 4772–4788 (2015).

50. Alain, C., Shen, D., Yu, H. & Grady, C. Dissociable memory- and response-related activity in parietal cortex during auditory spatial working memory. *Front. Psychol.* **1,** (2010).

51. Santoro, R. *et al.* Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 4799–4804 (2017).

52. Ladefoged, P. *Elements of Acoustic Phonetics*. *University of Chicago Press* (University of Chicago Press, 1962). doi:10.7208/chicago/9780226191010.001.0001

53. Bolt, R. H. *et al.* Identification of a Speaker by Speech Spectrograms. *Science (80-.).* **166,** 338–342 (1969).

54. Hayes, B. & Wilson, C. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguist. Inq.* **39,** 379–440 (2008).

55. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (International Conference on Learning Representations, ICLR, 2013).

56. Frank, S. L. & Willems, R. M. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Lang. Cogn. Neurosci.* **32,** 1192–1203 (2017).

57. Ding, N. & Simon, J. Z. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* **107,** 78–89 (2012).

58. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10,** (2016).

59. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37,** 6539–6557 (2017).

60. Hausfeld, L., Riecke, L., Valente, G. & Formisano, E. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *Neuroimage* **181,** 617–626 (2018).

61. Crosse, M. J., Butler, J. S. & Lalor, E. C. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* **35,** 14195–1420 (2015).

62. O'Sullivan, J. A. *et al.* Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb. Cortex* bht355 (2014).

63. Bednar, A. & Lalor, E. C. Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *Neuroimage* **205,** (2020).

64. Daube, C., Ince, R. A. A. & Gross, J. Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Curr. Biol.* **29,** 1924–1937.e9 (2019).

65. Guion, S. G., Flege, J. E., Akahane-Yamada, R. & Pruitt, J. C. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am.* **107,** 2711–2724 (2000).

66. Flege, J. E. The production and perception of foreign language speech sounds. *Hum. Commun. its Disord. a Rev. – 1988* 224–401 (1988).

67. Di Liberto, G. M. & Lalor, E. C. Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hear. Res.* **348,** 70–77 (2017).

68. Morgan-Short, K., Finger, I., Grey, S. & Ullman, M. T. Second Language Processing Shows

Increased Native-Like Neural Responses after Months of No Exposure. *PLoS One* **7,** e32974 (2012).

69. Cao, F., Tao, R., Liu, L., Perfetti, C. A. & Booth, J. R. High proficiency in a second language is characterized by greater involvement of the first language network: Evidence from Chinese learners of English. *J. Cogn. Neurosci.* **25,** 1649–1663 (2013).

70. Chang, C. B. The phonetics of second language learning and bilingualism. in *The Routledge Handbook of Phonetics* 427–447 (2019). doi:10.4324/9780429056253-16

71. Flege, J. E. Second Language Speech Learning: Theory, Findings, and Problems. *Speech Percept. Linguist. Exp. Issues Cross-Language Res.* 233–277 (1995). doi:10.1111/j.1600-0404.1995.tb01710.x

72. Brandmeyer, A., Farquhar, J. D. R., McQueen, J. M. & Desain, P. W. M. Decoding Speech Perception by Native and Non-Native Speakers Using Single-Trial Electrophysiological Data. *PLoS One* **8,** (2013).

73. Kimppa, L. *et al.* Acquisition of L2 morphology by adult language learners. *Cortex* **116,** 74–90 (2019).

74. Mueller, J. L. Electrophysiological correlates of second language processing. *Second Lang. Res.* **21,** 152–174 (2005).

75. Khachatryan, E., Camarrone, F., Fias, W. & Van Hulle, M. M. ERP Response Unveils Effect of Second Language Manipulation on First Language Processing. *PLoS One* **11,** e0167194 (2016).

76. Osterhout, L. *et al.* Second-language learning and changes in the brain. *J. Neurolinguistics* **21,** 509–521 (2008).

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

78. Brodbeck, C., Presacco, A., Anderson, S. & Simon, J. Z. Over-representation of speech in

older adults originates from early response in higher order auditory cortex. in *Acta Acustica united with Acustica* **104,** 774–777 (S. Hirzel Verlag GmbH, 2018).

79. Presacco, A., Simon, J. Z. & Anderson, S. Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. *PLoS One* **14,** e0213899 (2019).

80. Hjortkjaer, J., Märcher-Rørsted, J., Fuglsang, S. A. & Dau, T. Cortical oscillations and entrainment in speech processing during working memory load. *Eur. J. Neurosci.* (2018). doi:10.1111/ejn.13855

81. Kuhl, P. K. *et al.* Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences* **363,** 979–1000 (2008).

82. Best, C. T. & Tyler, M. D. Nonnative and second-language speech perception: Commonalities and complementarities. in *Language experience in second language speech learning: In honor of James Emil Flege* 13–34 (2007).

83. Chang, C. B. & Mishler, A. Evidence for language transfer leading to a perceptual advantage for non-native listeners. *J. Acoust. Soc. Am.* **132,** 2700–2710 (2012).

84. Flege, J. E. *The production of 'new' and 'similar' phones in a foreign language: evidence for the effect of equivalence classification*. *Journal of Phonetics* **15,** (1987).

85. Borovsky, A., Elman, J. L. & Kutas, M. Once is Enough: N400 Indexes Semantic Integration of Novel Word Meanings from a Single Exposure in Context. *Lang. Learn. Dev.* **8,** 278–302 (2012).

86. Kutas, M. & Hillyard, S. A. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science (80-.).* **207,** 203–205 (1980).

87. Chen, Q. *et al.* Automatic processing of taxonomic and thematic relations in semantic priming - Differentiation by early N400 and late frontal negativity. *Neuropsychologia* **64,** 54–62 (2014).

88. Manfredi, M., Cohn, N., De Araújo Andreoli, M. & Boggio, P. S. Listening beyond seeing:

Event-related potentials to audiovisual processing in visual narrative. *Brain Lang.* **185,** 1–8 (2018).

89. Wlotko, E. W. & Federmeier, K. D. So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuroimage* **62,** 356–366 (2012).

90. Coulson, S. & Kutas, M. Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neurosci. Lett.* **316,** 71–74 (2001).

91. Steinhauer, K. & Drury, J. E. On the early left-anterior negativity (ELAN) in syntax studies. *Brain Lang.* **120,** 135–162 (2012).

92. Meltzer, J. A. & Braun, A. R. P600-like positivity and left anterior negativity responses are elicited by semantic reversibility in nonanomalous sentences. *J. Neurolinguistics* **26,** 129–148 (2013).

93. Purves, D. *et al. Neuroscience, 2nd edition*. (de boeck, 2003).

94. Bialystok, E. & Hakuta, K. Confounded Age: Linguistic and Cognitive Factors in Age Differences for Second Language Acquisition. *Lawrence Erlbaum Assoc. Publ.* 161–181 (1999). doi:10.1017/S0272263101333053

95. Tucker, G. R. A Global Perspective on Bilingualism and Bilingual Education. *Socioling. Essent. Readings* **000,** 464–471 (2001).

96. Chang, C. B. Rapid and multifaceted effects of second-language learning on first-language speech production. *J. Phon.* **40,** 249–268 (2012).

97. Kartushina, N. & Frauenfelder, U. H. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Front. Psychol.* **5,** (2014).

98. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118,** 887–906 (2005).

99. Wang, K. & Shamma, S. Self-Normalization and Noise-Robustness in Early Auditory

Representations. *IEEE Trans. Speech Audio Process.* **2,** 421–435 (1994).

100. Lesenfants, D., Vanthornhout, J., Verschueren, E. & Francart, T. Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech. *J. Neural Eng.* (2019). doi:10.1088/1741-2552/ab3c92

101. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123,** 3878 (2008).

102. Boersma, P. & Weenink, D. Praat: doing phonetics by computer (Version 5.1. 05)[Computer program]. Retrieved May 1, 2009. (2009).

103. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (80-.).* **343,** 1006–1010 (2014).

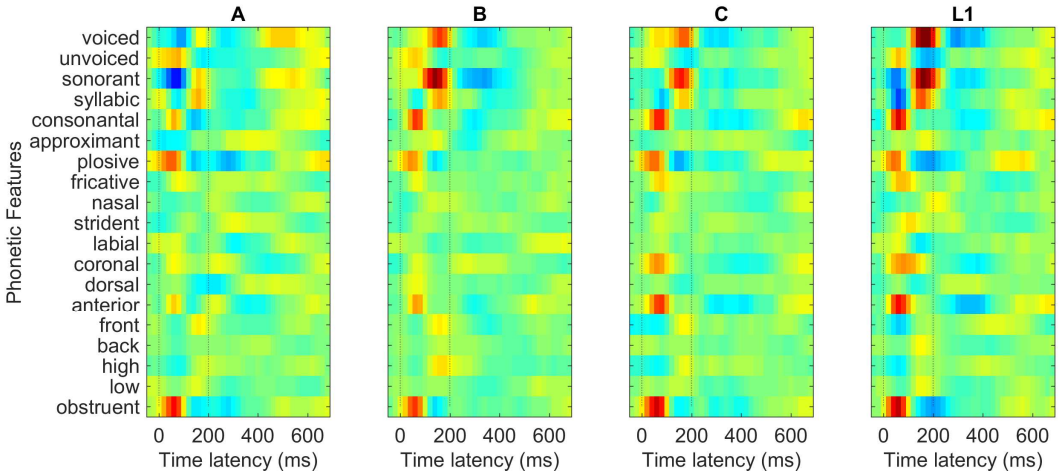
104. Goldwater, S. & Johnson, M. Learning OT Constraint Rankings Using a Maximum Entropy Model. *Proc. Stock. Work. Var. within Optim. Theory* 111–120 (2003).

105. Baroni, M., Dinu, G. & Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference* **1,** 238–247 (Association for Computational Linguistics (ACL), 2014).

106. Lalor, E. C., Power, A. J., Reilly, R. B. & Foxe, J. J. Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *J. Neurophysiol.* **102,** 349–359 (2009).

107. Ding, N., Chatterjee, M. & Simon, J. Z. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* **88,** 41–46 (2014).

108. Di Liberto, G. M. & Lalor, E. C. Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hear. Res.* **348,** 70–77 (2017).



Supplementary Figure 1. *Average temporal response function weights across subjects at the electrode Cz and peri-stimulus time-latencies from 0 to 600 ms for each proficiency level. These TRF weights were used to produce the result in Figure 3A.*

				Optimal decoding feature for	
		Feature	Description	L2 proficiency	L1 vs. L2
TRF weights		Envelope	Temporal dimension of the spectrogram response		
		Spectrogram	Frequency dimension of the spectrogram response		
		Phonetic features	Temporal and feature dimensions of the phoneme response		
		Phoneme onsets	Temporal dimension of the phoneme onset response		
		Phonotactics	Temporal and feature dimensions of the phoneme response		
		Semantic dissimilarity	All EEG channels and time-lags		
		Semantic dissimilarity Front-Back	Pz minus FPz for time-lags between 300 and 600 ms		
		Phonotactics short latencies	Phonotactics for time-lags between 80 and 140 ms		
		Phonotactics longer latencies	Phonotactics for time-lags between 300 and 700 ms		
EEG prediction correlations	SubjectSpecific	Env	Prediction r when using Env feature		
		Phn	Prediction r when using Phn feature		
		Pt	Prediction r when using Pt feature		
		Sd	Prediction r when using Sd feature		
		PhnEnv'Sgr - Env'Sgr	Increase in Prediction r when using Phn features		
	Generic	Envelope: generic A	Prediction r when using the Env generic model fit on the A group		
		Envelope: generic B	Prediction r when using the Env generic model fit on the B group		
		Envelope: generic C	Prediction r when using the Env generic model fit on the C group		
		Envelope: generic L1	Prediction r when using the Env generic model fit on the L1 group		
		PhnEnv'Sgr - Env'Sgr: generic A	Increase in prediction r when using Phn - generic model A group		
		PhnEnv'Sgr - Env'Sgr: generic B	Increase in prediction r when using Phn - generic model B group		
		PhnEnv'Sgr - Env'Sgr: generic C	Increase in prediction r when using Phn - generic model C group		
		PhnEnv'Sgr - Env'Sgr: generic L1	Increase in prediction r when using Phn - generic model L1 group		
		Sd: generic A	Prediction r when using the Sd generic model fit on the A group		
		Sd: generic B	Prediction r when using the Sd generic model fit on the B group		
		Sd: generic C	Prediction r when using the Sd generic model fit on the C group		
		Sd: generic L1	Prediction r when using the Sd generic model fit on the L1 group		

Supplementary Table 1. *The 26 features used for the proficiency decoding and L1 vs. L2 classification analyses. Features are grouped into TRF weights, EEG prediction correlations when using subject-specific models, and EEG prediction correlations when using generic models that were averaged within a particular proficiency group. A backward elimination procedure was used for feature selection. Features that were selected for the decoding are indicated with a colored cell.*