NeuroImage 192 (2019) 66–75

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Auditory and language contributions to neural encoding of speech features in noisy environments

Jiajie Zou [a,1], Jun Feng [b,c,d,1], Tianyong Xu [e], Peiqing Jin [a], Cheng Luo [a], Jianfeng Zhang [a], Xunyi Pan [f], Feiyan Chen [e], Jing Zheng [a], Nai Ding [a,*]

[a] Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang Univ., 310027, China
[b] Institute of Psychological Sciences, Hangzhou Normal University, Hangzhou, 311121, China
[c] Center for Cognition and Brain Disorders, Hangzhou Normal University, Hangzhou, 311121, China
[d] Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments, Hangzhou, 311121, China
[e] Bio-X Laboratory, Department of Physics, Zhejiang University, Hangzhou, 310027, China
[f] School of International Studies, Zhejiang University, Hangzhou, 310027, China

ARTICLE INFO

Keywords:
Electroencephalography
Speech envelope
Gain control
Neural entrainment

ABSTRACT

Recognizing speech in noisy environments is a challenging task that involves both auditory and language mechanisms. Previous studies have demonstrated human auditory cortex can reliably track the temporal envelope of speech in noisy environments, which provides a plausible neural basis for noise-robust speech recognition. The current study aimed at teasing apart auditory and language contributions to noise-robust envelope tracking by comparing the neural responses of 2 groups of listeners, i.e., native listeners and foreign listeners who did not understand the testing language. In the experiment, speech signals were mixed with spectrally matched stationary noise at 4 intensity levels and listeners' neural responses were recorded using electroencephalography (EEG). When the noise intensity increased, the neural response gain increased in both groups of listeners, demonstrating auditory gain control. Language comprehension generally reduced the response gain and envelope-tracking precision, and modulated the spatial and temporal profile of envelope-tracking activity. Based on the spatio-temporal dynamics of envelope-tracking activity, a linear classifier can jointly decode the 2 listener groups and 4 levels of noise intensity. Altogether, the results showed that without feedback from language processing, auditory mechanisms such as gain control can lead to a noise-robust speech representation. High-level language processing modulated the spatio-temporal profile of the neural representation of speech envelope, instead of generally enhancing the envelope representation.

1. Introduction

Speech perception is a complex process involving both auditory and language processing. Language processing can feed back and modulate basic auditory perception (Ganong, 1980; Warren, 1970). A sound feature that strongly contributes to speech intelligibility is the speech envelope, i.e., slow fluctuations (<16 Hz) in sound intensity (Drullman et al., 1994; Shannon et al., 1995). When listening to speech, neural activity tracking the speech envelope can be recorded either intracranially from the auditory cortex (Nourski et al., 2009) or non-invasively by magnetoencephalography/electroencephalography (MEG/EEG) (Ahissar et al., 2001; Ding and Simon, 2012b). In noisy environments, neural tracking of the speech envelope remains robust as long as the speech stream is attended to (Ding and Simon, 2012a; Kerlin et al., 2010; Mesgarani and Chang, 2012; O'Sullivan et al., 2014; Zion Golumbic et al., 2013). Although it is well established that cortical activity can track the speech envelope, it remains controversial whether speech-tracking activity is solely generated by general auditory

Abbreviations: MEG, magnetoencephalography; EEG, electroencephalography; SNR, signal-to-noise ratio; RMS, root mean square; LPC, linear predictive coding; EOG, electrooculogram; TRF, temporal response function; LDA, linear discriminant analysis; PCA, principal components analysis; SVD, singular value decomposition; FDR, false discovery rate; SEM, standard error of the mean; SD, standard deviation.

* Corresponding author. Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, 310027, China.
E-mail address: ding_nai@zju.edu.cn (N. Ding).
[1] These authors contribute equally.

https://doi.org/10.1016/j.neuroimage.2019.02.047
Received 26 July 2018; Received in revised form 31 January 2019; Accepted 19 February 2019
Available online 27 February 2019

mechanisms or also modulated by speech-specific neural computations (Ding and Simon, 2014).

One hypothesis is that envelope tracking responses are generated by non-speech-specific auditory mechanisms (Steinschneider et al., 2013), since the temporal envelope is a low-level acoustic feature well represented throughout the auditory system (Joris et al., 2004). Consistent with this hypothesis, envelope tracking responses can be seen in animals (David et al., 2009), and in humans listening to non-speech sounds, e.g., amplitude modulated noise or tones (Lalor et al., 2009; Wang et al., 2012). Some studies have found similar envelope tracking responses for intelligible speech and unintelligible speech such as time-reversed speech (Howard and Poeppel, 2010), and speech in an unknown language (Peña and Melloni, 2012). Based on the domain-general auditory encoding hypothesis, noise-robust neural tracking of the speech envelope can be explained by contrast gain control (Ding and Simon, 2013) or primitive auditory scene analysis, i.e., sound source segregation based on acoustic features (Bregman, 1990; Ding et al., 2014). Contrast gain control and primitive auditory scene analysis are general auditory mechanisms that have also been observed in primary auditory cortex of animals (Micheyl et al., 2005; Rabinowitz et al., 2011, 2013).

Another hypothesis assumes that envelope tracking responses reflect interactions between auditory and language processing. Consistent with this hypothesis, some studies have shown that when a speech signal is acoustically degraded to compromise intelligibility, neural tracking of the speech envelope shows reduced precision (Gross et al., 2013; Kong et al., 2015; Luo and Poeppel, 2007; Peelle et al., 2013). Furthermore, at the individual level, listeners showing more precise envelope tracking activities exhibit better speech understanding (Ding et al., 2014; Ding and Simon, 2013; Doelling et al., 2014). A potential concern about whether speech intelligibility directly modulates envelope tracking activity, however, is that intelligibility often covaries with other factors. Some of these factors are capable of modulating envelope tracking activity, such as acoustic changes, task difficulty, top-down attention, and individual hearing functions (Kayser et al., 2015; Lakatos et al., 2013; Petersen et al., 2017).

Here, we investigated how auditory and language mechanisms separately contribute to envelope-tracking speech responses in noisy environments. Behaviorally, it is known that language information increases speech intelligibility in noise (Miller et al., 1951). According to the domain-general auditory processing hypothesis, language knowledge facilitates speech recognition at a late stage, which is not reflected in the envelope tracking response. The interactive processing hypothesis, however, proposes that language processing feeds back and modulates envelope-tracking activity. To test these two hypotheses, we investigated the influence of language processing by comparing the neural tracking activity of two groups of listeners, i.e., native listeners of the testing language and foreign listeners who do not understand the testing language. A low-level auditory task, which did not require language comprehension, was employed to ensure attention. The speech signal was mixed with spectrally matched stationary noise at 4 signal-to-noise ratios (SNRs), and the envelope-tracking activity from both listener groups was recorded using EEG.

2. Materials and methods

2.1. Participants

Thirty-two adults participated in this experiment (18–29 years old; mean age, 22.9 years). All participants were right-handed and reported normal hearing, and were undergraduate or graduate students at Zhejiang University. Sixteen participants (8 females) were native Cantonese listeners while the other 16 participants (8 females) were native Mandarin listeners who had no previous knowledge of Cantonese. Mandarin is the only official language at school and Cantonese is not a dialect spoken in Zhejiang province. None of the native Mandarin listeners in the current experiment reported any exposure to Cantonese in their daily life. Participants received monetary payment for their participation and the experimental protocol was approved by the Institutional Review Board of the Zhejiang University Interdisciplinary Center for Social Sciences. Informed consents were obtained from all participants.

2.2. Stimuli and procedures

The speech recordings were selected from the audiobook *Legends of the Condor Heroes*, narrated in Cantonese by a male speaker. One hundred and sixty sections were randomly selected from the audiobook, with 15 s in duration for each section. All sections were normalized to the same intensity, measured by the root mean square (RMS). One hundred and twelve sections were used in normal trials, while the other forty-eight sections were used in outlier trials. Speech streams in normal trials were checked to ensure there were no adjacent repetition of sound segments. In outlier trials, a chunk of stimulus consisting of two syllables was randomly selected and immediately repeated. In half of the outlier trials, the two-syllable segment was repeated for one time, and in the other half the repetition was rendered twice. The boundaries of the two-syllable chunk were manually determined.

Spectrally matched stationary noise was generated using a 12-order linear predictive coding (LPC) model derived from all speech materials. The noise was mixed with speech at 4 different SNRs, i.e., +9 dB, −6 dB, −9 dB, and −12 dB. These 4 levels of SNRs were selected since a previous study has shown highly robust envelope-tracking neural responses with SNR above −6 dB (Ding and Simon, 2013). Here, the SNR of +9 dB was chosen as a baseline and the other 3 SNR levels were selected to characterize how the envelope-tracking activity was affected by SNR levels.

Forty trials were used in each SNR condition, including 28 normal trials and 12 outlier trials. Trials at different SNRs were mixed and presented in a randomized order. The intensity contrast of speech, i.e., the standard deviation (SD) of the envelope divided by its mean (Nelken et al., 1999), was reduced by the noise mixed in the speech, and the spectro-temporal features were distorted. The stimulus spectrogram and envelope, the intensity contrast of speech, and the correlation between the envelopes of the stimulus and the underlying clean speech are shown in Fig. 1 AB. After listening to each trial, listeners were expected to press a key on the keyboard to indicate the detection of repeated sound segments, key 0 for no repetition, key 1 for one repetition, and key 2 for double repetitions, respectively.

2.3. EEG recording and preprocessing

EEG responses were recorded using a 64-channel Biosemi ActiveTwo system, sampled at 2048 Hz. Two reference channels were placed at the left and right mastoids respectively and four channels were used to record horizontal and vertical electrooculogram (EOG). EEG signals were referenced offline by subtracting the average of the two mastoid recordings. EOG artifacts were regressed out using the least squares method (Ding et al., 2017). The EEG recordings and the speech envelopes were downsampled to 50 Hz and epoched based on the onset of every 15-s stimulus. A 1-s response starting from the stimulus-onset was removed to eliminate the onset response. Only responses to normal trials (112 trials for every listener) were used in analysis and EEG responses evoked by outliers, i.e., the trials with the repeated sound segments, were excluded.

2.4. EEG based envelope reconstruction

A linear decoder was used to reconstruct the temporal envelope of the underlying speech from the EEG response to the speech-noise mixture. The decoder applied a weighted average of the EEG response across time lags and channels using the following equation:

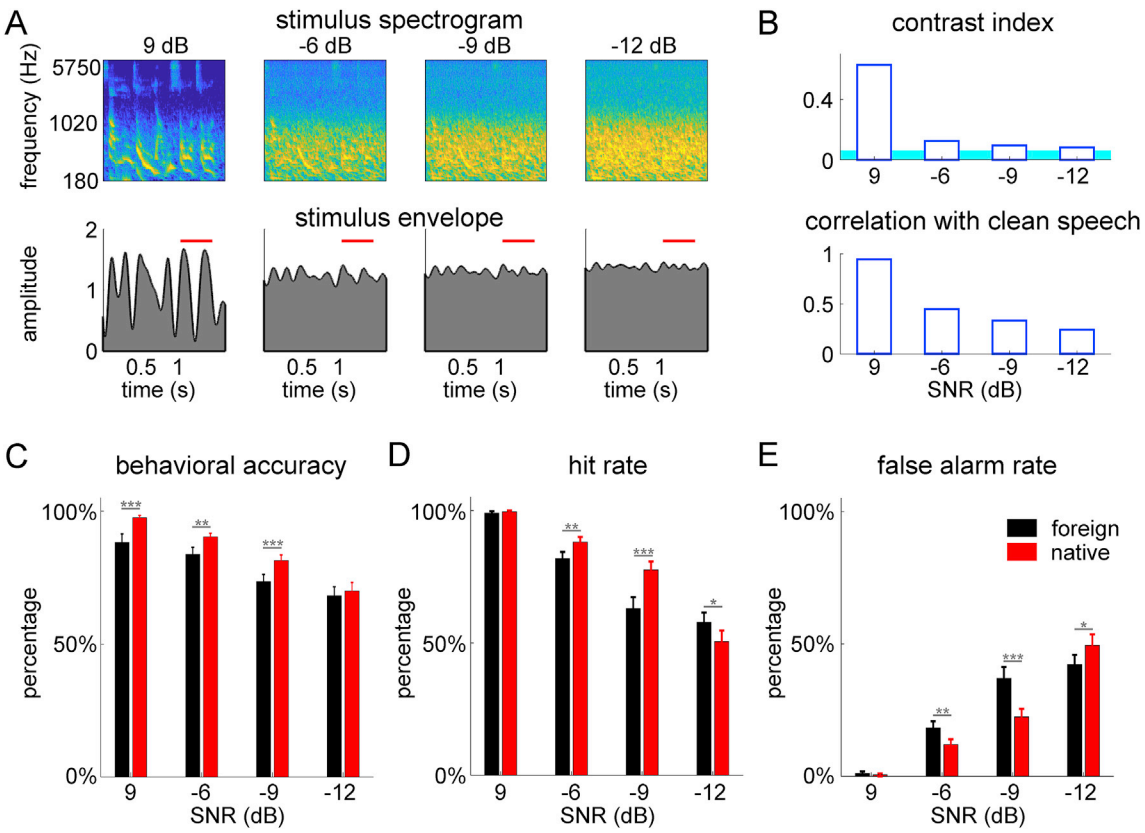

Fig. 1. Stimuli and behavioral results. (A) Stimulus spectrogram (top) and envelope (bottom) in the 4 SNR conditions. The stimuli shown here consist of a repeated stimulus segment, and the red lines illustrate the time interval when the segment repeats. (B) Contrast index of the stimulus (top) and the correlation between the stimulus envelope and the envelope of the underlying speech (bottom). The blue area covers the 95th percentile of the contrast index of stationary noise. (C) Percentage of trials correctly responded by native (red) and foreign listeners (black). All trials were used in this analysis. Native listeners exhibited significantly higher accuracy when SNR was above −12 dB. (DE) The hit rate and false alarm rate are shown in panel D and E respectively. Error bars represent 1 standard error of the mean (SEM) across listeners. Significant differences between native and foreign listeners are indicated by gray stars. *P < 0.05, **P < 0.01, ***P < 0.001 (bootstrap, FDR corrected).

$$\hat{s}(n) = \sum_{k=1}^{64} \sum_{m=1}^{M} D_k(m) r_k(n+m),$$

where $D_k(n)$ and $r_k(n)$ denoted the decoder weights and the EEG signals in channel k respectively. The order of the decoder, i.e., M, was 26, corresponding to a maximal time lag of 0.5 s. The decoder weights $D_k(n)$ were optimized so that the reconstructed envelope, $\hat{s}(n)$, approximated the underlying speech envelope. The decoder $D_k(n)$ was derived based on least-squares estimation with L2 regularization, i.e., normalized reverse correlation (Theunissen et al., 2001).

The neural reconstruction analysis was applied to every listener's dataset of neural responses individually. The reconstruction accuracy was measured by the Pearson correlation between the reconstructed envelope and the envelope of the underlying speech. It was evaluated using 10-fold cross validation: Each time 90% of data was used to train the decoder and the rest 10% of data was used to evaluate the reconstruction accuracy. The procedure was repeated 10 times and the 10 reconstruction accuracy values derived were averaged. The regularization parameter for neural reconstruction varied between 0 and 0.01, and 0.001 was determined as the optimal value as it yielded the highest neural reconstruction accuracy averaged across conditions and participants.

To characterize the frequency bands in which the EEG response was most correlated with the speech envelope, envelope reconstruction was performed separately for different bandpass filtered signals. In this analysis, a filter bank was used to decompose the speech envelope and the EEG responses into narrow bands. The filter bank contained 25 filters in a 1-Hz bandwidth, and the center frequencies increased in a linear manner from 1.5 Hz to 24.5 Hz in steps of 1 Hz. The frequency range in which the reconstruction accuracy was significantly higher than chance fell in the delta (1–4 Hz) and theta (4–8 Hz) bands (Fig. S1), consistent with the literature (Ding and Simon, 2012b; Luo and Poeppel, 2007). Therefore subsequent analyses were restricted to 1–8 Hz.

A sigmoid function was used to characterize the relationship between behavioral accuracy and neural reconstruction accuracy, denoted as y and x respectively in the following. The sigmoid function, i.e., $y = A + (1 - A)/(1 + exp(-\alpha(x - m)))$, included 3 parameters, i.e., A, α, and m, which referred to the lower asymptote, the growth rate, and the location of this sigmoid function, respectively. The 3 parameters were fitted using the least squares method.

2.5. Temporal response function

Neural reconstruction characterized how accurately the speech or stimulus envelope was represented in the brain by integrating EEG responses across time and channels. A neural encoding model, i.e., the temporal response function (TRF), was used to further characterize the spatial and temporal patterns of the neural responses. The TRF was formulated as the following:

$$r(n) = \sum_{m=1}^{T} TRF(m) s(n - m + m_1) + e(n),$$

where $r(n)$, $TRF(n)$, $s(n)$, and $e(n)$ denoted the EEG response, the TRF, the sound envelope, and the residual error respectively. The order of the TRF

model, i.e., T, was 26, corresponding to 0.5 s. The m_1 was −5, so that the TRF contained a 5-sample, i.e., 0.1 s, prestimulus interval. The TRF was computed using the least squares estimation with L2 regularization. The regularization parameter was tuned to provide the highest predictive power, which was measured by the correlation between the predicted neural responses, i.e., $\sum_{m=1}^{T} TRF(m)s(n-m+m_1)$, and the actual responses. Specifically, the regularization parameter was varied between 0 and 0.3. Results showed that the optimal parameters were 0.18, 0.18, and 0.14 for the actual stimulus, the normalized stimulus, and the clean speech respectively, as they yielded the highest predictive power averaged across all channels, conditions, and participants.

2.6. Classification of listeners and SNR levels

A classification analysis was performed to quantify whether the spatial and temporal profile of the TRF was modulated by language processing and stimulus SNR. In this analysis, the 2 listener groups, i.e., native and foreign listeners, and the 4 SNR levels were classified based on the TRF predictive power or the spatial and temporal profile of the TRF. First, the dimension reduction was performed for 1-dimensional (1×64) TRF predictive power by applying the linear discriminant analysis (LDA), which projected high-dimensional data to dimensions in which data from different classes were maximally separated (Duda et al., 2012). Then, a Euclidean distance-based classifier was used to classify listener groups and SNR levels based on the first 2 LDA dimensions. The classifier determined the center of each class based on the training set. For each sample in the testing set, the distance to each class center was calculated and its class was determined based on the nearest center. For instance, if a testing sample was closer to the center of class A than any other centers in Euclidean distance, it was assigned to class A. Since there were 16 listeners in each group, an 8-fold cross validation was used to evaluate the performance of the classifier: Each time, data of 14 listeners were used to train the classifier while those of the other 2 listeners were used to evaluate the accuracy of classification.

When the LDA was applied to the 2-dimensional (time by channel) TRF ($T \times 64$), the TRF was reshaped to a 1-dimensional vector ($64T \times 1$) by concatenating the responses from different channels. As the number of features was expected to be smaller than the number of samples in the LDA, the principal components analysis (PCA) was performed before the LDA to reduce the $64T \times 1$ vector to a relative lower dimension whose number was determined to have a highest classification accuracy. As the TRF was reshaped to a $64T \times 1$ vector, the LDA feature vector was also a $64T \times 1$ vector. To characterize the spatial and temporal profile of the LDA feature vector, it was reshaped back to a 2-dimensional (time by channel) matrix and the spatial and temporal profile was the first left and right singular vector extracted by singular value decomposition (SVD).

2.7. Statistical tests

The bootstrap significance test is a bias-corrected and accelerated procedure (Efron and Tibshirani, 1994). In this procedure, the data of all participants were resampled 5000 times with replacement, and each time the data sampled were averaged across participants, therefore a total of 5000 mean values were produced. For paired comparisons (e.g., comparisons between SNR conditions), bootstrap was performed to test the differences between conditions; if N_S out of the 5000 mean values were greater (or smaller) than 0, the significance level was N_S/5000. For unpaired comparisons, i.e., comparisons between listener groups, data from the foreign listeners were resampled and used to estimate the null distribution. If the mean value across native listeners was greater (or smaller) than N_S out of the 5000 resampled mean values of the foreign listeners, the significance level was N_S/5000.

In the neural reconstruction analysis, chance-level reconstruction accuracy was estimated by constructing surrogate neural responses. A surrogate neural response was created by circularly shifting the actual response by time lag between 100 s and 300 s in steps of 2 s. This procedure resulted in 101 surrogate neural responses. If the actual reconstruction accuracy exceeded the 95% percentile of the chance-level reconstruction accuracy, it was considered statistically significant ($P < 0.05$). A similar process was operated to estimate the chance-level predictive power and RMS of the TRF. When multiple comparisons were performed, the p-value was further adjusted using the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

3. Results

3.1. Behavioral results

In the experiment, listeners were instructed to detect repeated sound segments in a continuous speech stream. The percentage of correctly responded trials is shown in Fig. 1C for native and foreign listeners. A 2-way repeated-measures ANOVA (listener group × SNR) revealed significant main effects for both factors: SNR ($F_{3,90} = 73.17$, $P < 0.001$) and listener group ($F_{1,30} = 4.66$, $P = 0.039$). The interaction between the two factors, however, was not significant ($F_{3,90} = 1.77$, $P < 0.158$). Additionally, the behavioral accuracy was significantly higher for native listeners at 9 dB ($P < 0.001$, bootstrap, FDR corrected), −6 dB ($P = 0.001$, bootstrap, FDR corrected) and −9 dB ($P < 0.001$, bootstrap, FDR corrected). Considering a possible response bias in the behavioral results and the difficulty in calculating D-prime for a 3-alternative forced choice experiment, the hit rate and the false alarm rate were presented separately (Fig. 1DE). It was found that as the noise level increased, the hit rate decreased and the false alarm rate increased. In addition, native listeners showed a significantly higher hit rate and a significantly lower false alarm rate at −6 dB (hit rate: $P = 0.001$, and false alarm rate: $P < 0.001$, bootstrap, FDR corrected) and −9 dB (hit rate: $P = 0.003$, and false alarm rate: $P < 0.001$, bootstrap, FDR corrected).

3.2. Neural reconstruction of speech

To study whether the speech was reliably represented in the brain against background noise, the temporal envelope of the underlying speech was reconstructed based on the neural responses to the speech-noise mixture (Fig. 2A). In other words, the analysis focused on how precisely the speech envelope was "picked out" from the speech-noise mixture and encoded in the brain. Neural reconstruction accuracy monotonically decreased with decreasing SNR for both native and foreign listeners. A 2-way repeated-measures ANOVA (listener group × SNR) was performed on the data of reconstruction accuracy, which showed a significant main effect of SNR ($F_{3,90} = 136.41$, $P < 0.001$) and a significant interaction between the two factors ($F_{3,90} = 5.08$, $P = 0.003$). The main effect of the listener group was nearly significant ($F_{1,30} = 4.05$, $P = 0.053$). At +9 dB and −6 dB SNRs, foreign listeners showed higher reconstruction accuracy than native listeners (+9 dB: $P < 0.001$, and −6 dB: $P < 0.001$, bootstrap, FDR corrected). When the relationship between reconstruction accuracy and SNR was fitted by a line, the data from foreign listeners produced a significantly steeper slope than those from the native group ($P < 0.001$, bootstrap). This result suggested that the neural reconstruction accuracy exhibited greater sensitivity to noise in foreign listeners than in native listeners.

The relationship between grand averaged behavioral accuracy and neural reconstruction accuracy is shown in Fig. 2B and was fitted by a sigmoid function. The growth rate, which determines the slope of the sigmoid function, was significantly higher for native than for foreign listeners ($P < 0.001$, bootstrap, FDR corrected). The fitted location parameter of the sigmoid function was not significantly different between listener groups ($P = 0.294$, bootstrap, FDR corrected). The relationship between behavioral accuracy and reconstruction accuracy for individual listeners in every SNR condition is shown in Fig. 2C. The correlation between these two measurements was not significantly

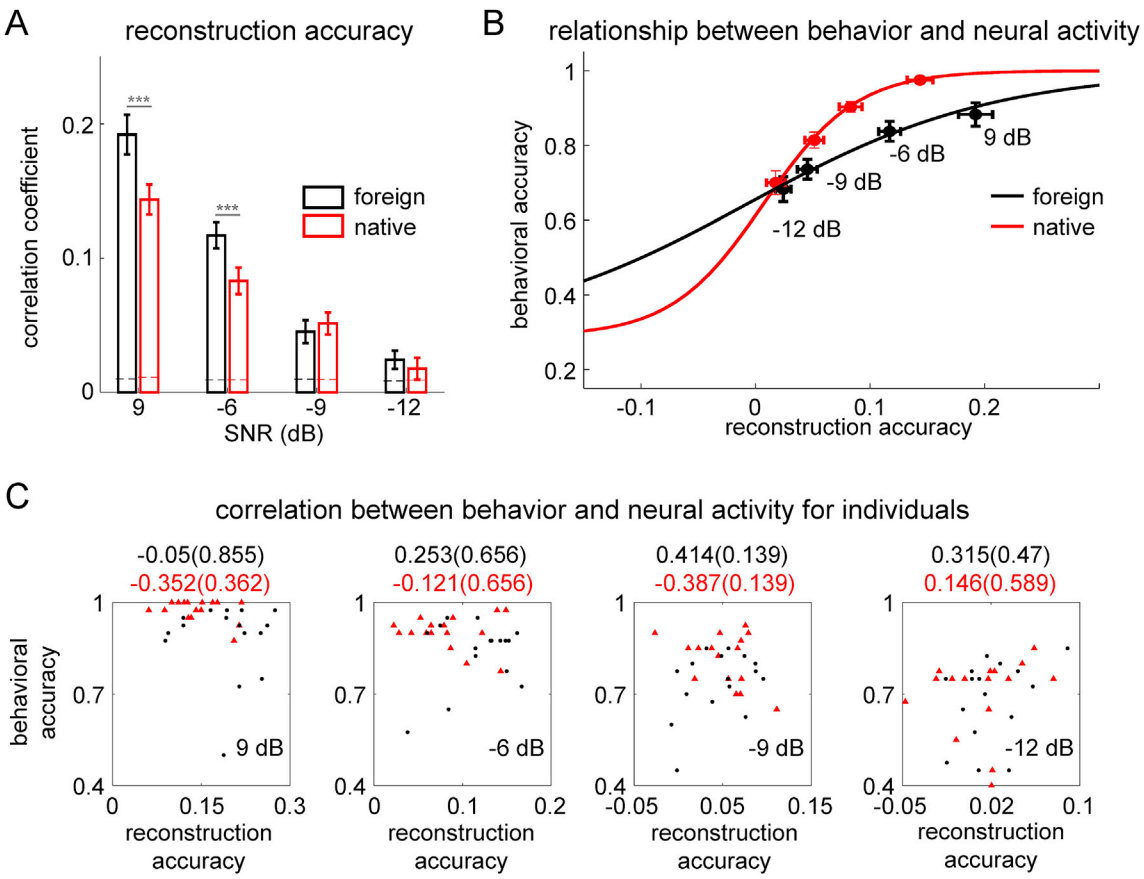

Fig. 2. Accuracy of neural reconstruction of speech envelope and its relation to behavioral results. (A) Reconstruction accuracy for native (red) and foreign listeners (black) in every SNR condition. Dashed lines denote the chance level. Significant differences between native and foreign listeners are indicated by gray stars. ***P < 0.001 (bootstrap, FDR corrected). (B) Relationship between behavioral accuracy and neural reconstruction accuracy. Error bars represent 1 SEM across listeners. The relationship between neural and behavioral accuracy was fitted by a sigmoid function for each listener group. (C) Correlation between behavioral accuracy and neural reconstruction accuracy for individual subjects. Each red triangle denotes data from 1 native listener, while each black dot denotes data from 1 foreign listener. The R-value (P-value) for native (red) and foreign (black) listeners are shown in the title of each plot. No significant correlation between behavioral and reconstruction accuracy was observed in any of the 4 SNR conditions.

different from 0 in any condition ($P > 0.05$).

3.3. Temporal response function (TRF)

The decoding analysis characterized the precision of neural encoding of the speech envelope. In the next analysis, the temporal and spatial profile of envelope tracking activity was characterized by a TRF analysis. The TRF estimated the time course of neural activity evoked by a unit power increase in the stimulus (Ding and Simon, 2012b). First, the TRF was derived from the actual stimuli presented in trials, i.e., the speech-noise mixture, to characterize the encoding of the sound input. The TRF predictive power described how precisely the speech envelope was tracked, similar to the reconstruction accuracy, but it was calculated for every single EEG channel instead of a combination of all channels. In general, the reconstruction accuracy and the predictive power averaged across 64 channels were affected by SNR and the listener group in a similar pattern (Fig. S2). At −9 dB SNR, however, the channel-averaged predictive power was significantly higher for native than foreign listeners ($P = 0.033$, bootstrap, FDR corrected).

When the SNR was above −12 dB, the channel-average predictive power was greater than chance ($P < 0.05$) (Fig. S2), and the topography of TRF predictive power generally showed a centro-frontal distribution (Fig. 3A). To test if the topographic distribution showed fine differences between conditions, a classification analysis was performed to distinguish the 2 listener groups and 4 SNR levels on the basis of the topography of predictive power. In this analysis, the predictive power from 64 channels was reduced to 2 dimensions using LDA. The 2 listener groups were well separated at high SNRs (Fig. 3B). The LDA decomposed the 64-channel topography into spatial patterns, and the spatial patterns corresponding to the first two LDA dimensions are shown in Fig. 3C. When classifying the 64-channel predictive power into 8 categories (2 listener groups x 4 SNR levels), the classification accuracy reached 38.3%, significantly higher than the chance-level performance of 12.5% (binomial test, $P < 0.001$). The confusion matrix of the classification results is presented in Fig. 3D, which shows the decoded categories of data from each group of listeners in every SNR condition.

The amplitude and time course of the TRF were analyzed as follows. The RMS of the TRF across channels (normalized by subtracting the pre-stimulus value) is shown in Fig. 4AB. A 2-way repeated-measures ANOVA (listener group × SNR) revealed a main effect of SNR on the TRF total power ($F_{3,90} = 10.16$, $P < 0.001$). Neither the main effect of listener group ($F_{1,30} = 0.71$, $P = 0.406$) nor the interaction between listener group and SNR ($F_{3,90} = 2.14$, $P = 0.101$) was significant. The TRF total power significantly increased from 9 dB to −6 dB (foreign: $P < 0.001$, and native: $P = 0.01$, bootstrap, FDR corrected). For foreign listeners, the response gain decreased from −6 dB to −9 dB ($P = 0.003$, bootstrap, FDR corrected); for native listeners, the response gain decreased from −9 dB to −12 dB ($P = 0.01$, bootstrap, FDR corrected). These results demonstrated that the neural response gain exhibited a U-shaped relationship with the stimulus SNR: Relatively low-intensity noise induced an increase of neural response gain to compensate the loss of stimulus contrast. The increase of neural response gain, however, stopped against high-intensity

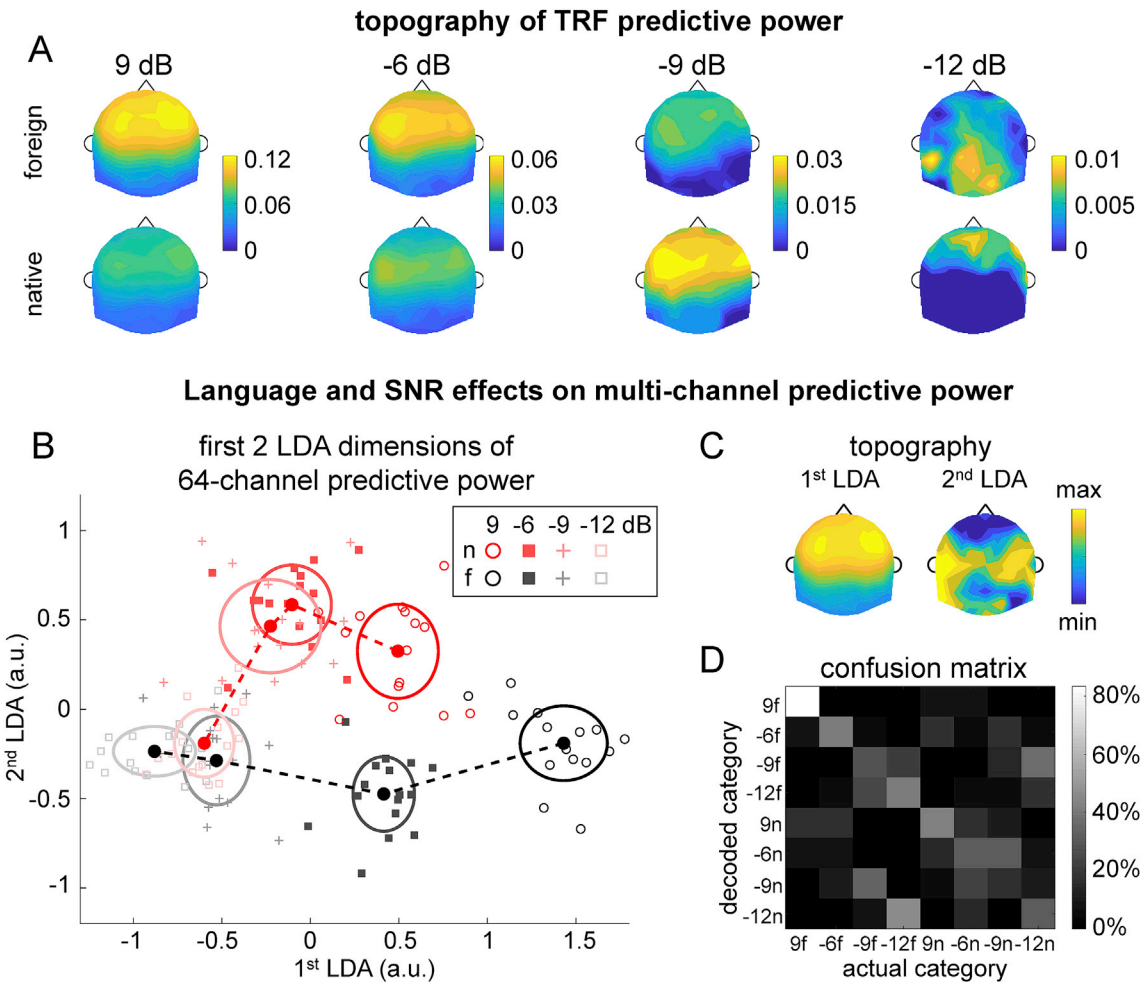

Fig. 3. Topography of TRF predictive power differs across different listener groups and SNR levels. (A) Topography of predictive power for foreign (upper) and native (lower) listeners. (B) Scatterplot of the first 2 LDA dimensions of the 64-channel predictive power. Lighter colors denote lower SNRs. Each marker represents data from 1 listener. The center of each ellipse is the mean across listeners while the radius represents 1 SD across listeners. (C) Topography of the first 2 LDA dimensions. (D) Confusion matrix of the classification accuracy shows a diagonal structure, indicating above-chance performance of classification (f: foreign listeners; n: native listeners).

noise exposure.

The previous analysis showed that the response gain was modulated by the stimulus SNR. In the following, we characterized how the observed response gain change deviated from an ideal model in which the change in stimulus contrast was fully compensated by the neural response gain. In this analysis, a new TRF was derived based on the amplitude-normalized actual stimulus (the speech-noise mixture) envelope, i.e., the actual stimulus envelope divided by its standard deviation representing an ideal contrast gain control. If the change in neural response gain fully compensated the contrast reduction in stimulus, the TRF derived from the amplitude-normalized stimulus was expected to show an SNR-invariant pattern. The results, however, showed a monotonic decrease of TRF total power as the SNR decreased, demonstrating that gain control was not ideal at any SNR level.

In the following, a classification analysis was performed to test if language knowledge and stimulus SNR modulated the spatial and temporal profile of TRF (Fig. S3). Using a procedure similar to that applied to the classification of TRF predictive power, listener groups and SNR levels can be jointly classified with an accuracy of 45.3% (higher than chance, binomial test, $P < 0.001$). Since the predictive power analysis has already revealed the modulation of spatial properties of the TRF by both language knowledge and SNR, what remained unclear was whether the temporal information alone sufficed to classify listener groups and stimulus SNRs. To address this issue, the spatial dimension of the TRF was removed by averaging TRF across channels. The subsequent analysis yielded a classification accuracy above chance (43%, binomial test, $P < 0.001$), demonstrating that language knowledge and stimulus SNR modulated the TRF time course. In the classification analysis, the TRF model was derived from stimuli with normalized amplitude. For the TRF model derived from the actual stimulus, similar classification results were obtained based on the multi-channel TRF and the channel-averaged TRF (43% and 41.4% respectively, significantly above chance, binomial test, $P < 0.001$).

To further compare whether the actual stimulus envelope or the underlying speech envelope better modeled the EEG responses, another TRF was derived based on the envelope of the underlying clean speech (Fig. 4E). Fig. 4G compares the predictive power of two models, one derived from the speech-noise mixture and the other derived from the underlying clean speech. The two models exhibited similar predictive power for data from native listeners. For foreign listeners, however, the model derived from the clean speech yielded slightly better predictive power at +9 dB ($P < 0.001$, bootstrap, FDR corrected) and −6 dB ($P = 0.038$, bootstrap, FDR corrected).

The power of the TRF derived from the clean speech, as shown in Fig. 4E, monotonically decreased with decreasing SNR. For the TRF total power, a 2-way repeated-measures ANOVA (listener group × SNR) revealed a significant main effect of SNR ($F_{3,90} = 34.85$, $P < 0.001$) and a significant interaction between listener group and SNR ($F_{3,90} = 2.99$,

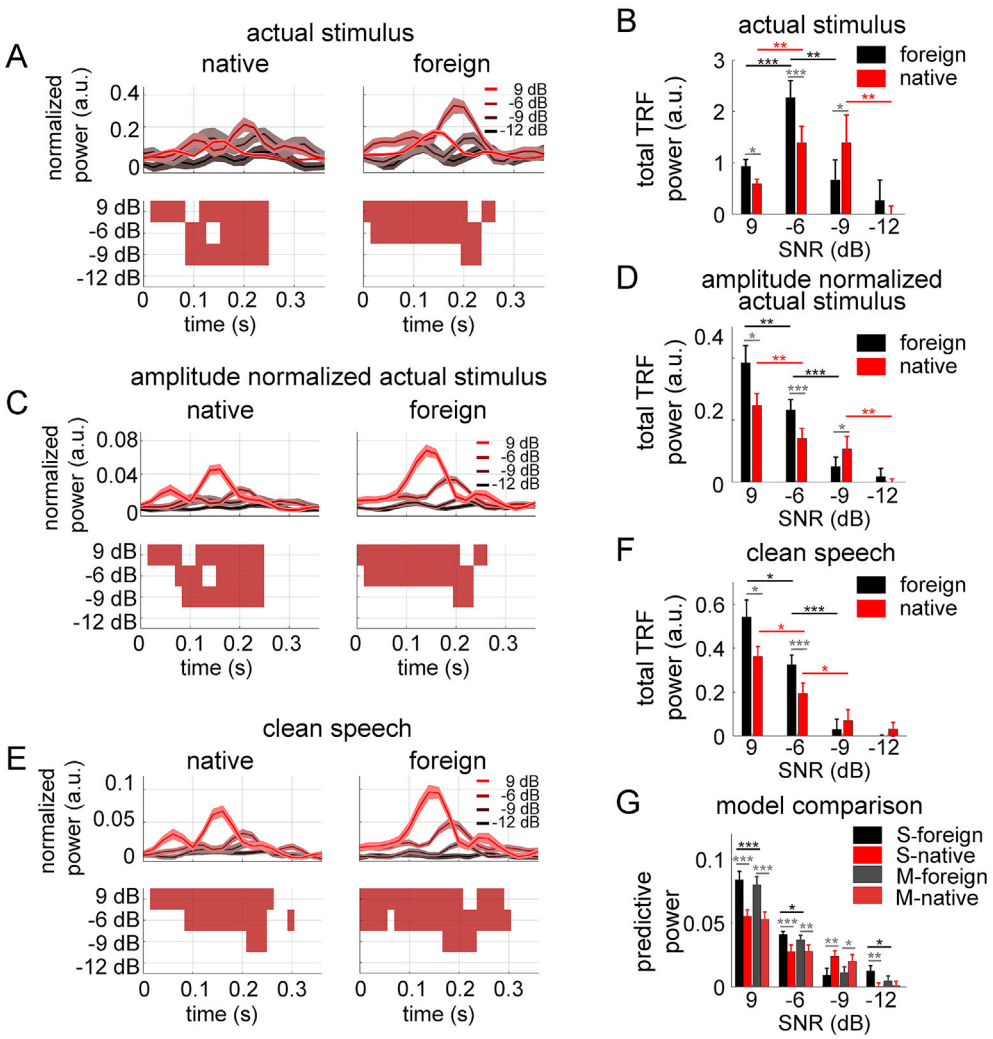

Fig. 4. TRF derived from the actual stimulus (AB), the amplitude-normalized actual stimulus (CD), and the underlying clean speech (EF). (ACE) The upper panel shows the TRF time course, i.e., the RMS across channels, and the shaded area denotes 1 SEM across listeners. Red bricks in the lower panel represent the time intervals in which the TRF amplitude is significantly higher than chance, and data from all SNR conditions are stacked vertically. (BDF) Total power of the TRF. (G) Predictive power of the TRF derived from the underlying clean speech (S), and the speech-noise mixture (M). Error bars represent 1 SEM across listeners. Significant differences between SNR conditions and between models are indicated by stars: red for native listeners; black for foreign listeners. Significant differences between native and foreign listeners are indicated by gray stars. *P < 0.05, **P < 0.01, ***P < 0.001, (bootstrap, FDR corrected).

P = 0.035), but no significant main effect of listener group ($F_{1,30} = 2.15$, P = 0.153) (Fig. 4F). In addition, the neural responses of foreign listeners presented a significantly higher normalized power than those of the native listeners at +9 dB (P = 0.034, bootstrap, FDR corrected) and −6 dB (P = 0.007, bootstrap, FDR corrected).

4. Discussion

This study demonstrates that the cortical representation of speech envelope was modulated by both auditory and language mechanisms, when listeners performed a low-level auditory task that did not require speech comprehension. Envelope-tracking cortical activity could be generated by domain-general auditory mechanisms, and auditory gain control mechanisms could lead to a noise-robust representation even for unintelligible speech. Language processing, however, modulated the precision, and also the spatial and temporal profile of the envelope-tracking cortical activity.

4.1. Auditory mechanisms underlying noise-robust speech responses

When a speech signal was embedded in noise, the dynamic range of its intensity fluctuations, i.e., the intensity contrast, was compressed (Fig. 1B). If the auditory cortex passively followed the stimulus envelope, the amplitude of the envelope-tracking response was expected to fall when the noise level increased. Nevertheless, a number of studies have shown that a reduction in intensity contrast can be compensated at various auditory processing stages (Dean et al., 2005; Robinson and McAlpine, 2009), so that in auditory cortex the neural activity is only weakly affected by the intensity contrast of the stimulus (Rabinowitz et al., 2011). MEG studies in humans show that neural tracking of the speech envelope is robust to noise and is barely influenced by the intensity contrast of the stimulus when the SNR is above ~ -6 dB and the speech intelligibility is above ~50% (Ding and Simon, 2013). Robust neural encoding of the speech envelope is attributed to auditory gain control mechanisms. In the current study, it was demonstrated that such mechanisms can be applied to unintelligible speech with no need for feedback from high-level language processing (Fig. 4AB). In addition, our results indicate that, compared with the actual stimulus, the underlying clean speech showed comparable or better predictive power for both native and foreign listeners (Fig. 4G). These results suggest that even when listening to an unknown language the brain can actively filter out background noise and selectively encode the speech signal.

4.2. Language processing contributes to speech processing in noise

Behaviorally, it has long been demonstrated that language processing facilitates speech recognition in noisy environments. High-context sentences can be better understood in noisy environments than low-context sentences or random words (Miller et al., 1951). Furthermore, prior knowledge of sentence content can facilitate listeners' understanding of sentence in noise (Helfer and Freyman, 2005; Jones and Freyman, 2012; Kidd et al., 2014; Yang et al., 2007). In the present study, compared with

foreign listeners, native listeners showed better performance (i.e., higher behavioral accuracy, higher hit rate, and lower false alarm rate) in detection of repeated sound segments when the stimuli were presented at −6 dB and −9 dB (Fig. 1CDE), which demonstrated that language knowledge could facilitate a low-level auditory task. The behavioral studies cannot, however, directly demonstrate whether prior information, either linguistic knowledge or context, modulates low-level auditory representations of speech, or only influences late language processing.

4.3. Speech intelligibility and cortical envelope tracking

Whether speech intelligibility affects cortical tracking of the speech envelope has been extensively studied. The basic logic is to manipulate speech intelligibility and to investigate the related changes in the cortical responses to the speech envelope. In general, speech intelligibility can be manipulated in 3 different ways, i.e., by changing the acoustic properties of speech, by changing the linguistic context, and by observing the responses from listeners with different levels of speech recognition ability. The relationship between speech intelligibility and cortical envelope tracking is complicated. The current study and a number of previous studies found that language processing, which occurs only for intelligible speech, suppresses envelope tracking. In contrast, a large number of other studies showed better envelope tracking for more intelligible speech. This apparent discrepancy is discussed in the following. It is argued that the spatio-temporal properties of envelope-tracking activity and the types of intelligibility manipulations need to be taken into consideration when discussing the relationship between intelligibility and cortical envelope tracking. Domain-general auditory processing, which is largely independent from higher-level language processing, is likely to occur in lower-level auditory cortex. This kind of auditory responses are sensitive to acoustic manipulations of speech. Interactions between auditory and language processing occur in higher-level cortical areas and the interactions between the two factors could be highly heterogeneous.

4.3.1. Acoustic manipulations

A number of studies have investigated whether envelope-tracking cortical activity deteriorates when speech intelligibility is compromised by acoustic manipulations. Nevertheless, inconsistent results have been obtained, especially in quiet listening environments. For example, some studies have reported no overall reduction in the precision of envelope-tracking activity when the speech is played backward (Howard and Poeppel, 2010; Zoefel and VanRullen, 2016), when the speech fine structure is corrupted (Ding et al., 2014), or when the speech is time-compressed (Nourski et al., 2009). Nevertheless, other studies have found the attenuation in envelope-tracking activity when the speech is played backward (Di Liberto et al., 2015; Gross et al., 2013), when the spectro-temporal fine structure is corrupted (Luo and Poeppel, 2007; Peelle et al., 2013), when the speech envelope is corrupted (Doelling et al., 2014), or when the speech is time-compressed (Ahissar et al., 2001).

The inconsistency of the results is probably attributed to the complexity of speech processing pathway: Research on both humans and animals has shown that a faithful representation of the sound input in the auditory periphery is gradually transformed into a representation that is largely invariant to the listening background in auditory cortex (Ding and Simon, 2013; Power et al., 2012; Rabinowitz et al., 2013; Schneider and Woolley, 2013). Furthermore, functional imaging studies have shown that different cortical areas show different levels of sensitivity to acoustic manipulations of speech (Davis and Johnsrude, 2003; Overath et al., 2015; Scott et al., 2004), manipulations of the linguistic context (Friederici et al., 2000), and language proficiency (Kim et al., 1997). A recent intracranial EEG study has also suggested that different cortical areas are differentially affected by stimulus manipulation, e.g., time compression (Davidesco et al., 2018). Therefore spatial information needs to be taken into consideration when discussing how envelope-tracking activity is related to speech intelligibility: Lower-level auditory cortical areas may track the sound envelope of speech and non-speech sound in a similar manner, while higher-level language-related areas may track the speech envelope of intelligible and unintelligible speech in different patterns.

In noisy listening environments, research findings have been consistent, reporting that when an acoustic interference signal reduces speech intelligibility, it also reduces the precision of cortical envelope tracking (Ding and Simon, 2013; Kong et al., 2015; Vanthornhout et al., 2018). Furthermore, trial-by-trial variations in envelope tracking activity can predict trial-by-trial variations in listeners' performance in a speech comprehension task (Keitel et al., 2018). The current results also show a decrease in envelope tracking accuracy when the noise level increases. A related method to manipulate speech intelligibility is to use the priming paradigm. In some conditions, an unintelligible sentence becomes intelligible if the listeners hear an intelligible version of the same sentence in advance. Using the priming approach, however, two studies have concluded that intelligibility does not modulate envelope tracking (Baltzell et al., 2017; Millman et al., 2015). Nevertheless, the priming paradigm cannot distinguish sensory-level adaptation effect caused by priming and the intelligibility variation caused by priming.

4.3.2. Linguistic manipulations

As to the manipulation of the linguistic content of speech, one study observes more accurate delta-band envelope tracking for sentences constructed by pseudowords, compared with sentences constructed by real words (Mai et al., 2016). One explanation for this phenomenon is that linguistic context provides additional cues for speech perception and reduces the demand for processing of acoustic cues in the speech envelope. Another explanation is that neural tracking of the linguistic content (Ding et al., 2016) competes with neural tracking of envelope and therefore reduces envelope-tracking activity. In the current study, these explanations are also potential reasons for the observation of lower envelope tracking precision in native listeners in contrast to better behavioral performance achieved (Fig. 2B & S2): It is possible that native listeners tend to pay less attention to the speech envelope since they have access to higher-level linguistic cues. Similar phenomenon has also been found in the statistical learning paradigm: When listeners learn that syllables are grouped into artificial words, neural tracking of the syllabic-rate speech envelope tends to reduce (Buiatti et al., 2009).

4.3.3. Individual differences

The relationship between envelope tracking activity and intelligibility can also be characterized by individual differences. One study has compared the neural responses of native Italian and Spanish listeners when they listen to Italian, Japanese, and Spanish utterances (Peña and Melloni, 2012). It is shown that speech intelligibility does not change neural responses in low-frequency power, a gross measure that can reflect both envelope-tracking activity and other neural activity. Using the ERP approach, it has also been shown that the responses to consonant clusters (Wagner et al., 2012) and vowels (Näätänen et al., 1997) can be significantly affected by listeners' native language. The current study extends these previous studies by explicitly analyzing neural tracking of the speech envelope. In another study, the strength of the envelope-tracking response is comparable for native and foreign listeners when they listen to sentences constructed by isochronously presented syllables (Ding et al., 2016). The current study, however, finds more precise neural tracking of the speech envelope in foreign listeners for utterances with a natural rhythm.

Even within the population of young native listeners, individuals vary in their ability to understand speech in noisy environments. Studies have consistently shown that in challenging listening conditions, individuals showing more precise envelope-tracking activity tend to understand speeches better (Ding et al., 2014; Ding and Simon, 2013; Doelling et al., 2014; Kong et al., 2015; Vanthornhout et al., 2018). In the current study, however, the precision of envelope-tracking activity was not correlated with how well listeners could detect repeated sound segments (Fig. 2C).

A possible reason was that the repetition-detection task was designed to maintain a basic level of attention, rather than to comprehensively assess intelligibility. The detection performance only depended on speech audibility and the listener's attentional state at the moment when the repetition occurs. Given the rarity of sound segments repetition (at most once within a 15-s stimulus), this task may not reflect how clearly the listeners can hear speech in general.

Within the population of native listeners, when aging is considered, more precise neural tracking of the speech envelope is observed in older listeners compared with young listeners, in both quiet and noisy listening environments (Decruy et al., 2018; Presacco et al., 2016), even though older listeners have more trouble understanding speech in noise.

Taking all the studies into consideration, it is suggested that speech-tracking activity is generally related to intelligibility at the individual level within a relatively homogeneous population, e.g., young native listeners. It remains unclear, however, whether this effect is driven by individual differences in auditory ability or language ability. Across population groups, the correlated dynamic between intelligibility and envelope tracking precision seems to be heterogeneous.

4.4. Technical consideration

In EEG studies, results can be affected by the choice of reference. Therefore we compared the results obtained by using two common references, i.e., the linked mastoids reference and the average reference that averaged all 64 channels as a reference (Fig. S4). It was found that neural reconstruction accuracy (Fig. S4A) was barely affected by the choice of reference. The TRF predictive power was lower at +9 dB SNR when the average reference was adopted (Fig. S4B, native listeners: $P = 0.005$, and foreign listeners: $P = 0.023$, bootstrap, FDR corrected), but the topography still showed a centro-frontal distribution (Fig. S4C). These results demonstrated that the reconstruction accuracies derived from the use of two different references, i.e., the linked mastoid and the average of all channels, were similar. With linked mastoids as the reference, the TRF predictive power showed a lower spatial resolution but was higher in average.

5. Summary

In sum, the current results and previous studies (Buiatti et al., 2009; Mai et al., 2016; Presacco et al., 2016) demonstrated that when acoustic properties of the stimuli are controlled, language processing generally reduces the precision of envelope-tracking activity in quiet listening environments. Critically, we also found that the precision of envelope tracking activity was differentially modulated by speech intelligibility in different EEG channels, so the topography of the TRF predictive power could be used for classifying listener groups (Fig. 3). Similarly, the time course of the TRF, i.e., the RMS across channels, also sufficed to the classification of listener groups (Fig. S3). These results suggest that future research need to consider the spatio-temporal dynamics of neural activity when discussing how intelligibility modulates envelope-tracking responses.

Declarations of interest

None.

Acknowledgement

Work supported by National Natural Science Foundation of China 31500873 (ND), 31771248 (ND), Zhejiang Provincial Natural Science Foundation of China LR16C090002 (ND), and research funding from the State Key Laboratory of Industrial Control Technology, Zhejiang University (ND). We thank Yuhan Lu and Lin Chen for helpful comments on earlier versions of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.02.047.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. U. S. A. 98, 13367–13372.
Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. J. Neurophysiol. 118, 3144–3151.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 289–300.
Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. The MIT Press, Cambridge.
Buiatti, M., Pena, M., Dehaene-Lambertz, G., 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. Neuroimage 44, 509–519.
David, S.V., Mesgarani, N., Fritz, J.B., Shamma, S.A., 2009. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. 29, 3374–3386.
Davidesco, I., Thesen, T., Honey, C.J., Melloni, L., Doyle, W., Devinsky, O., Ghitza, O., Schroeder, C., Poeppel, D., Hasson, U., 2018. Electrocorticographic Responses to Time-Compressed Speech Vary across the Cortical Auditory Hierarchy. bioRxiv, p. 354464.
Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.
Dean, I., Harper, N.S., McAlpine, D., 2005. Neural population coding of sound level adapts to stimulus statistics. Nat. Neurosci. 8, 1684–1689.
Decruy, L., Vanthornhout, J., Francart, T., 2018. Evidence for Enhanced Neural Tracking of the Speech Envelope Underlying Age-Related Speech-In-Noise Difficulties. bioRxiv, p. 489237.
Di Liberto, Giovanni, M., O'Sullivan, James A., Lalor, Edmund C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. 25, 2457–2465.
Ding, N., Chatterjee, M., Simon, J.Z., 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88, 41–46.
Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., Poeppel, D., 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front. Hum. Neurosci. 11, 481.
Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164.
Ding, N., Simon, J.Z., 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. U. S. A. 109, 11854.
Ding, N., Simon, J.Z., 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89.
Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33, 5728–5735.
Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. 8, 311.
Doelling, K., Arnal, L., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85, 761–768.
Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95, 1053–1064.
Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.
Friederici, A.D., Meyer, M., von Cramon, D.Y., 2000. Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang. 74, 289–300.
Ganong, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6, 110–125.
Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11, e1001752.
Helfer, K.S., Freyman, R.L., 2005. The role of visual speech cues in reducing energetic and informational masking. J. Acoust. Soc. Am. 117, 842–849.
Howard, M.F., Poeppel, D., 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J. Neurophysiol. 104, 2500–2511.
Jones, J.A., Freyman, R.L., 2012. Effect of priming on energetic and informational masking in a same-different task. Ear Hear. 33, 124–133.
Joris, P.X., Schreiner, C.E., Rees, A., 2004. Neural processing of amplitude-modulated sounds. Physiol. Rev. 84, 541–577.
Kayser, S.J., Ince, R.A., Gross, J., Kayser, C., 2015. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. J. Neurosci. 35, 14691–14701.
Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol. 16, e2004473.
Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. 30, 620–628.

Kidd Jr., G., Mason, C.R., Best, V., 2014. The role of syntax in maintaining the integrity of streams of speech. J. Acoust. Soc. Am. 135, 766–777.
Kim, K.H., Relkin, N.R., Lee, K.-M., Hirsch, J., 1997. Distinct cortical areas associated with native and second languages. Nature 388, 171.
Kong, Y.-Y., Somarowthu, A., Ding, N., 2015. Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. J. Assoc. Res. Otolaryngol. 16, 783–796.
Lakatos, P., Schroeder, C.E., Leitman, D.I., Javitt, D.C., 2013. Predictive suppression of cortical excitability and its deficit in schizophrenia. J. Neurosci. 33, 11692–11702.
Lalor, E.C., Power, A.J., Reilly, R.B., Foxe, J.J., 2009. Resolving precise temporal processing properties of the auditory system using continuous stimuli. J. Neurophysiol. 102, 349–359.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.
Mai, G., Minett, J.W., Wang, W.S., 2016. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. Neuroimage 133, 516–528.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.
Micheyl, C., Tian, B., Carlyon, R.P., Rauschecker, J.P., 2005. Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48, 139–148.
Miller, G.A., Heise, G.A., Lichten, W., 1951. The intelligibility of speech as a function of the context of the test materials. J. Exp. Psychol. 41, 329–335.
Millman, R.E., Johnson, S.R., Prendergast, G., 2015. The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. J. Cognit. Neurosci. 27, 533–545.
Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432.
Nelken, I., Rotman, Y., Yosef, O.B., 1999. Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397, 154.
Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Matthew, A., Howard, I., Brugge, J.F., 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. 29, 15564–15574.
O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2014. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebr. Cortex 25, 1697–1706.
Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. 18, 903–911.
Peña, M., Melloni, L., 2012. Brain oscillations during spoken sentence processing. J. Cognit. Neurosci. 24, 1149–1164.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebr. Cortex 23, 1378–1387.
Petersen, E.B., Wostmann, M., Obleser, J., Lunner, T., 2017. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J. Neurophysiol. 117, 18–27.
Power, A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., Lalor, E.C., 2012. At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. 35, 1497–1503.
Presacco, A., Simon, J.Z., Anderson, S., 2016. Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. J. Neurophysiol. 116, 2346–2355.
Rabinowitz, N.C., Willmore, B.D., King, A.J., Schnupp, J.W., 2013. Constructing noise-invariant representations of sound in the auditory pathway. PLoS Biol. 11, e1001710.
Rabinowitz, N.C., Willmore, B.D., Schnupp, J.W., King, A.J., 2011. Contrast gain control in auditory cortex. Neuron 70, 1178–1191.
Robinson, B.L., McAlpine, D., 2009. Gain control mechanisms in the auditory pathway. Curr. Opin. Neurobiol. 19, 402–407.
Schneider, D.M., Woolley, S.M., 2013. Sparse and background-invariant coding of vocalizations in auditory scenes. Neuron 79, 141–152.
Scott, S.K., Rosen, S., Wickham, L., Wise, R.J.S., 2004. A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. J. Acoust. Soc. Am. 115, 813–821.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.
Steinschneider, M., Nourski, K.V., Fishman, Y.I., 2013. Representation of speech in human auditory cortex: is it special? Hear. Res. 305, 57–73.
Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., Gallant, J.L., 2001. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Netw. Comput. Neural Syst. 12, 289–316.
Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech intelligibility predicted from neural entrainment of the speech envelope. J. Assoc. Res. Otolaryngol. 19, 181–191.
Wagner, M., Shafer, V.L., Martin, B., Steinschneider, M., 2012. The phonotactic influence on the perception of a consonant cluster/pt/by native English and native Polish listeners: a behavioral and event related potential (ERP) study. Brain Lang. 123, 30–41.
Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., Simon, J.Z., 2012. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. J. Neurophysiol. 107, 2033–2041.
Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.
Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B.A., Li, L., 2007. The effect of voice cuing on releasing Chinese speech from informational masking. Speech Commun. 49, 892–904.
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77, 980–991.
Zoefel, B., VanRullen, R., 2016. EEG oscillations entrain their phase to high-level features of speech sound. Neuroimage 124, 16–23.