

## Two Stages of Speech Envelope Tracking in Human Auditory Cortex

### Modulated by Speech Intelligibility

Na Xu<sup>1,7</sup>, Baotian Zhao<sup>2</sup>, Lu Luo<sup>4</sup>, Kai Zhang<sup>2</sup>, Xiaoqiu Shao<sup>1</sup>, Guoming Luan<sup>5,8</sup>, Qian

Wang<sup>5,6,9\*</sup>, Wenhan Hu<sup>3\*</sup>, Qun Wang<sup>1,7,8\*</sup>

<sup>1</sup>Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China

<sup>2</sup>Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China

<sup>3</sup>Beijing Neurosurgical Institute, Capital Medical University, Beijing 100070, China

<sup>4</sup>School of Psychology, Beijing Sport University, Beijing 100084, China

<sup>5</sup>Beijing Key Laboratory of Epilepsy, Epilepsy Center, Sanbo Brain Hospital, Capital Medical University, Beijing 100093, China

<sup>6</sup>School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China

<sup>7</sup>National Clinical Research Center for Neurological Diseases, Beijing 100070, China

<sup>8</sup>Beijing Institute of Brain Disorders, Collaborative Innovation Center for Brain Disorders, Capital Medical University, Beijing 100069, China

<sup>9</sup>IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

#### \*Correspondence:

Qun Wang

E-mail: wangq@ccmu.edu.cn

Wenhan Hu

E-mail: huwenhan88@163.com

Qian Wang

E-mail: wangqianpsy@pku.edu.cn

## Abstract

The envelope is essential for speech perception. Recent studies have shown that cortical activity can track the acoustic envelope. However, whether the tracking strength reflects the extent of speech intelligibility processing remains controversial. Here, using stereo-electroencephalogram (sEEG) technology, we directly recorded the activity in human auditory cortex while subjects listened to either natural or noise-vocoded speech. These two stimuli have approximately identical envelopes, but the noise-vocoded speech does not have speech intelligibility. We found two stages of envelope tracking in auditory cortex: an early high- $\gamma$  (60-140 Hz) power stage (delay  $\approx$  49 ms) that preferred the noise-vocoded speech, and a late  $\theta$  (4-8 Hz) phase stage (delay  $\approx$  178 ms) that preferred the natural speech. Furthermore, the decoding performance of high- $\gamma$  power was better in primary auditory cortex than in non-primary auditory cortex, consistent with its short tracking delay. We also found distinct lateralization effects: high- $\gamma$  power envelope tracking dominated left auditory cortex, while  $\theta$  phase showed better decoding performance in right auditory cortex. In sum, we suggested a functional dissociation between high- $\gamma$  power and  $\theta$  phase: the former reflects fast and automatic processing of brief acoustic features, while the latter correlates to slow build-up processing facilitated by speech intelligibility.

**KEYWORDS:** stereo-electroencephalogram (sEEG); envelope tracking; high- $\gamma$  band power;  $\theta$  band phase; spectrally degraded speech

## INTRODUCTION

The temporal envelopes of sounds contain critical information for speech understanding (Shannon et al., 1995; Smith et al., 2002). When listening to speech, cortical activities, especially those in auditory cortex, can temporally track the acoustic envelopes (Poeppel and Assaneo, 2020). The strength of cortical envelope tracking may reflect the extent of speech processing (Di Liberto et al., 2018; Etard and Reichenbach, 2019; Vanthornhout et al., 2018). However, acoustic signals with no speech intelligibility (i.e., time-reversed speech or music) could also be well tracked by the cortical activities (Doelling and Poeppel, 2015; Harding et al., 2019; Howard and Poeppel, 2010). Therefore, whether cortical envelope tracking functionally contributed to speech intelligibility remains unclear.

To address this long-standing issue, an idea is to test the cortical activities to sounds with identical temporal envelopes but different intelligibility. Following this line of thought, a noise-vocoding technique can degrade the spectral regularity while preserving the speech envelope (Davis et al., 2005). Moreover, varying the number of frequency bands of the vocoder modulates the speech intelligibility effectively: with decreasing number of frequency bands, the intelligibility of the noise-vocoded speech was gradually destroyed (Davis and Johnsrude, 2003). However, how intelligibility level modulates cortical tracking of noise-vocoded speech remained puzzled. Some previous studies showed that cortical envelope tracking strength enhanced with increased speech intelligibility (Ding et al., 2014; Peelle et al., 2013), while others found not (Millman et al., 2015; Rimmele et al., 2015). In a recent study, Hauswald and collages (2020) found an inverted U-shaped relationship between the cortical envelope tracking strength and speech intelligibility.

These inconsistent findings implied the existence of different neural mechanisms of speech envelope tracking in auditory cortex, which has been rarely addressed. Electrophysiological evidence showed that both low and high-frequency cortical activities could track speech envelope (Gourevitch et al., 2020). In low-frequency bands, one of the most robust observations was that envelope could be tracked by the phase in  $\theta$  band (4-8 Hz). On the other hand, the high-frequency envelope tracking was mainly contributed by the fluctuated power in high- $\gamma$  band ( $\sim$ 70-150 Hz) (Kubanek et al., 2013; Kulasingham et al., 2020; Nourski et al., 2009; Synigal et al., 2020; Zion Golumbic et al., 2013). Both  $\theta$  phase envelope tracking and high- $\gamma$  power could be modulated by noise-vocoded speech in different intelligibility levels (Ding et al., 2014; Hauswald et al., 2020; Nourski et al., 2019; Peelle et al., 2013). However, there is still a lack of direct comparison between  $\theta$  phase and high- $\gamma$  power tracking patterns in auditory cortex, hindering our understanding of their roles in representing speech intelligibility information.

Using stereo-electroencephalogram (sEEG) technology in ten patients with refractory epilepsy, we directly recorded the electrophysiological signals in human auditory cortex when patients were passively listening to a natural or a two-band noise-vocoded speech. In the single-contact level, we analyzed the envelope tracking strength, tracking delay, and decoding accuracy of both high- $\gamma$  (60-140 Hz) power and  $\theta$  phase (4-8 Hz) in Heschl's gyrus (HG) and superior temporal gyrus (STG). HG includes mostly the primary auditory cortex, and STG is considered the non-primary auditory cortex (Clarke and Morosan, 2012). We then compared the tracking pattern and decoding performance in different hemispheres and brain areas. Our results revealed distinct processing stages of speech intelligibility subserved by high- $\gamma$  power and  $\theta$  phase in auditory cortex.

## RESULTS

### Auditory responsive contacts

We first selected the sEEG contacts in Heschl's gyrus (HG) and superior temporal gyrus (STG) in all ten patients. Out of 115 contacts, 75 contacts (HG,  $n = 20$ ; STG,  $n = 55$ ) (Figure 2A) with both significant auditory evoked high- $\gamma$  signal and broadband signal were further analyzed.

----- Insert Figure 2 about here -----

### Early high- $\gamma$ power envelope tracking preferred the noise-vocoded speech

The patients passively listened to a natural speech and a noise-vocoded speech (duration = 3.275 s), each with a repetition of 30 times. In each sEEG contact, we analyzed the envelope tracking strength ( $\rho$ ) and delay ( $\tau$ ) by conducting a cross-correlation test between time series of neural high- $\gamma$  power and acoustic envelope (Figure 2B).

Figure 2C showed waveforms of the envelopes and high- $\gamma$  power under natural and noise-vocoded speech conditions from one representative contact. As shown in Figure 2D, we estimated the optimal tracking strength and tracking delay under each speech condition at the single-contact level, where a positive lag time indicates that high- $\gamma$  power lags the speech envelope. The envelope tracking strength peaked with a 30 ms optimal tracking delay for natural speech and a 50 ms optimal tracking delay for noise-vocoded speech (natural  $\rho = 0.18$ ; noise vocoded  $\rho = 0.48$ ; Figure 2D). We tested the statistical significance of high- $\gamma$  power envelope tracking strength for each contact. Seventy-two contacts (out of 75 contacts) showed significant

tracking and were kept for further comparisons. We compared both the optimal tracking delay and strength of high- $\gamma$  power under two speech conditions using paired *t*-tests. The tracking delay for natural speech was significantly shorter than that for noise-vocoded speech (natural:  $49.2 \pm 4.3$  ms, mean  $\pm$  SE; noise-vocoded:  $61.4 \pm 4.0$  ms;  $t_{71} = -2.146$ ,  $p = 0.035$ ), while the tracking strength for natural speech condition was significantly lower than that for noise-vocoded speech (natural:  $0.22 \pm 0.02$ ; noise-vocoded:  $0.35 \pm 0.02$ ;  $t_{71} = 10.909$ ,  $p < 0.001$ ; Figure 2E and F). These results revealed that the high- $\gamma$  power tracked the speech envelope in the early stage of auditory processing (< 100 ms), and the tracking strength was enhanced when speech intelligibility was destroyed.

### **Late $\theta$ phase envelope tracking preferred the natural speech**

We also analyzed the envelope tracking strength ( $\rho$ ) and delay ( $\tau$ ) between neural  $\theta$  phase and  $\theta$  band-filtered envelope phase along the time series using the cerebro-acoustic coherence (CACoh) index (Harding et al., 2019; Peelle et al., 2013). Figure 2G showed waveforms of the  $\theta$  band-filtered envelope and the neural  $\theta$  signal under natural and noise-vocoded speech conditions, extracted from the same contact in Figure 2C. Similar to high- $\gamma$  power tracking, we estimated the peak tracking strength and optimal tracking delay of  $\theta$  phase under each speech condition at the single-contact level, where a positive lag time indicates that  $\theta$  phase lags the speech envelope (Figure 2H). For each contact, we built a null distribution for  $\theta$  phase envelope tracking by permutation (100 times) to examine its significance. Forty-six contacts (out of 75 contacts) were significant and included in the further analysis. We compared the optimal tracking delay and strength of the  $\theta$  phase under two speech conditions using paired *t*-tests. We

found that the  $\theta$  phase tracking delay for natural speech was significantly shorter than that for noise-vocoded speech (natural:  $178.1 \pm 5.0$  ms; noise-vocoded:  $198.2 \pm 3.1$  ms;  $t_{45} = -3.903$ ,  $p < 0.001$ ), while the tracking strength for natural speech was significantly larger than that for noise-vocoded speech (natural:  $0.38 \pm 0.05$ ; noise-vocoded:  $0.14 \pm 0.03$ ;  $t_{45} = 5.882$ ,  $p < 0.001$ ; Figure 2I and J). These results revealed that the  $\theta$  phase tracked the speech envelope in the late stage of auditory processing ( $> 100$  ms), and the tracking strength decreased when speech intelligibility was destroyed.

### **Left hemisphere lateralization of high- $\gamma$ power envelope tracking**

Previous findings suggest that high-frequency and low-frequency neural oscillations in the human auditory cortex showed functional lateralization in hemispheres (Giroud et al., 2020; Morillon et al., 2012; Poeppel, 2003) and functional dominance in brain areas (Fontolan et al., 2014). Thus, we investigated whether the envelope tracking strength of high- $\gamma$  power and  $\theta$  phase also showed lateralization or brain area effects. We pooled the data across natural and noise-vocoded speech conditions, and test the difference between hemisphere/brain areas using  $t$ -tests. Contacts with strong high- $\gamma$  power tracking strength clustered in the left hemisphere (Figure 3A), and  $t$ -tests results confirmed that the envelope tracking strength of high- $\gamma$  power of the left hemisphere was significantly larger than that of the right hemisphere (left:  $0.36 \pm 0.02$ ; right:  $0.20 \pm 0.01$ ;  $t_{142} = 6.450$ ,  $p < 0.001$ ) (Figure 3B). However, we did not observe significant difference for high- $\gamma$  power envelope tracking between brain areas (HG:  $0.27 \pm 0.03$ ; STG:  $0.28 \pm 0.02$ ;  $t_{142} = -0.550$ ,  $p = 0.583$ ) (Figure 3B). These results suggested a left hemisphere lateralization for high- $\gamma$  power envelope tracking.

Figure 3C showed the  $\theta$  phase envelope tracking strength of single contacts.  $\theta$  phase envelope tracking between was also compared between hemisphere/brain areas, and the results of *t*-tests showed no significant difference in HG and STG (HG:  $0.19 \pm 0.04$ ; STG:  $0.29 \pm 0.04$ ;  $t_{90} = -1.523$ ,  $p = 0.131$ ), and no significant difference between left and right hemispheres (left:  $0.23 \pm 0.04$ ; right:  $0.30 \pm 0.04$ ;  $t_{90} = -1.039$ ,  $p = 0.301$ ) (Figure 3D).

----- Insert Figure 3 about here -----

### **Distinct decoding performance of high- $\gamma$ power and the $\theta$ phase**

By far, our results suggest that high- $\gamma$  power and  $\theta$  phase tracked the envelopes of natural and noise-vocoded speech with different tracking strengths and delays. To examine whether these tracking patterns reflected the acoustic information representation in auditory cortex, we decoded the speech condition with single-trial high- $\gamma$  power or the  $\theta$  phase for each contact using the support vector machine (SVM) approach via leave-one-out cross-validations.

For high- $\gamma$  power, speech condition could be decoded in 49% of the contacts (statistical significance above the chance level). Chi-square tests showed that the HG had higher proportions of significant decoded contacts than STG (HG: 80.0%; STG: 38.2%;  $\chi^2 = 10.261$ ,  $p = 0.001$ ), while no significance was found between left and right hemisphere (left: 50.0%; right: 48.6%;  $\chi^2 = 0.014$ ,  $p = 0.907$ ; Figure 4B). We further testes the decoding accuracy between different brain areas and different hemispheres. The results revealed that high- $\gamma$  power decoding accuracy in HG contacts were significantly higher than that in STG contacts (HG:  $0.76 \pm 0.03$ ;

STG:  $0.63 \pm 0.02$ ;  $t_{73} = 3.270$ ,  $p = 0.002$ ), but no significance was found between left and right hemisphere (left:  $0.68 \pm 0.02$ ; right:  $0.65 \pm 0.03$ ;  $t_{73} = 0.657$ ,  $p = 0.513$ ; Figure 4C). These results indicated that high- $\gamma$  power decoding performance was predominant in Heschls' gyrus.

----- Insert Figure 4 about here -----

For  $\theta$  phase, speech condition could be decoded in 89% of the contacts. Chi-square tests showed that the right hemisphere had higher proportions of significant decoded contacts than the left hemisphere (left: 81.6%; right: 97.3%;  $\chi^2 = 4.861$ ,  $p = 0.027$ ), while no significance was found between HG and STG (HG: 85.0%; STG: 90.9%;  $\chi^2 = 0.537$ ,  $p = 0.463$ ; Figure 4E). We then tested the decoding accuracy between different brain areas and different hemispheres. The results revealed that  $\theta$  phase decoding accuracy in right hemisphere were significantly higher than that in left hemisphere (left:  $0.76 \pm 0.02$ ; right:  $0.86 \pm 0.02$ ;  $t_{73} = -3.842$ ,  $p < 0.001$ ), but no significance was found between HG and STG (HG:  $0.81 \pm 0.03$ ; STG:  $0.81 \pm 0.02$ ;  $t_{73} = -0.104$ ,  $p = 0.917$ ; Figure 4F). These results suggested a right hemisphere dominance of  $\theta$  phase decoding performance.

Finally, we compared the decoding performance of high- $\gamma$  power versus  $\theta$  phase and found that not only the proportion of significant contacts was larger when using  $\theta$  phase ( $\chi^2 = 28.219$ ,  $p < 0.001$ ), but also the mean decoding accuracy using  $\theta$  phase was better than that using high- $\gamma$  power ( $\theta$  phase:  $0.81 \pm 0.01$ ; high- $\gamma$  power:  $0.67 \pm 0.02$ ;  $t_{74} = 7.706$ ,  $p < 0.001$ ; Figure 4G).

## DISCUSSION

Using sEEG recordings, we revealed two distinct processing stages of speech envelope in human auditory cortex modulated by speech intelligibility. In the early stage, high- $\gamma$  power tracked speech envelope with a short latency ( $\approx 49$  ms) and enhanced when the speech sound was incomprehensible. In the late stage,  $\theta$  phase tracked speech envelope with a long latency ( $\approx 178$  ms) and decreased with degrading speech intelligibility. Consistent with the tracking latency, we also found that the spectra-temporal information in high- $\gamma$  power represented better in primary auditory cortex than in non-primary auditory cortex. Furthermore, we observed the different hemisphere lateralization of high- $\gamma$  power and  $\theta$  phase, suggesting that these two neural activities may have highly specific linguistic significance.

### **Early high- $\gamma$ power and late $\theta$ phase envelope tracking**

We showed reliable envelope tracking in high- $\gamma$  power and  $\theta$  phase differed in their tracking delay. The relative short tracking delay (about 50 ms) of high- $\gamma$  power is in line with prior work (Nourski et al., 2014; Zion Golumbic et al., 2013). In contrast,  $\theta$  phase tracking delay is long ( $>100$  ms), which is also consistent with previous studies (Di Liberto et al., 2015; Harding et al., 2019; Zou et al., 2021). And the finding that both of high- $\gamma$  power and  $\theta$  phase showed the shorter tracking delay for the natural speech compared with that for the noise-vocoded speech, suggests the sensitivity to natural human speech (Yu and Wang, 2021). Sound encoding is a hierarchical, feedforward processing framework in human auditory cortex (Hickok and Poeppel, 2007). The distinct tracking delays of high- $\gamma$  power and  $\theta$  phase indicate they may be involved in different stages of speech processing. The tracking delay in high- $\gamma$  power is similar to the

evoked onset response of human auditory cortex (Elhilali et al., 2004). Moreover, we also revealed a better decoding performance using high- $\gamma$  power feature clustered in Heschl's gyrus, which suggests the high- $\gamma$  power envelope tracking is an early response in primary auditory cortex. We did not observe a significant effect for brain areas for  $\theta$  phase decoding performance (Figure 4). However, we found a trend of larger  $\theta$  phase tracking strength in STG than Heschl's gyrus (Figure 3).

### **Distinct functional roles of high- $\gamma$ power and $\theta$ phase**

It has been suggested that there may be different functional roles for neural activities at different frequencies in speech processing (Mai et al., 2016). A previous ECoG study also showed a distinction between high- $\gamma$  power and low-frequency in tracking the envelope of time-compressed speech in human auditory cortex (Nourski et al., 2009). Specifically, high- $\gamma$  power, but not low-frequency, tracked the envelope of speech even when it was not comprehensible. As indicated in previous studies, high- $\gamma$  power is susceptible to the regularity of auditory events (Eliades et al., 2014; Yellamsetty and Bidelman, 2018), which may reflect the neural activity more related to the acoustic feature. In contrast,  $\theta$  phase envelope tracking was modulated by speech intelligibility to a large extent (Ding et al., 2014; Hauswald et al., 2020; Peelle et al., 2013). Our finding that high- $\gamma$  power envelope tracking preferred the noise-vocoded speech, while  $\theta$  phase envelope tracking preferred the natural speech (Figure 2), supports that these two bands may reflect distinct speech-specific processing.

Our data further showed that compared with high- $\gamma$  power,  $\theta$  phase feature decoding yielded a larger proportion of significant contacts and better decoding accuracy (Figure 3), indicating  $\theta$

phase may reflect the speech processing related to the intelligibility. Our findings align well with a recent study (Synigal et al., 2020), showing that the envelope of the attended stimuli could be better reconstructed using low-frequency (1-15 Hz) compared that using high- $\gamma$  power. Besides, there exists a relationship between the functions at different frequencies (Canolty et al., 2006; Wang et al., 2008; Zhang et al., 2020). Combining high- $\gamma$  power and low-frequency feature could better reconstruct the speech envelope (Akbari et al., 2019; Synigal et al., 2020) and optimize selective neuronal tracking of the envelope of the attended speaker (Zion Golumbic et al., 2013). Thus, high- $\gamma$  power and  $\theta$  phase envelope tracking may play different roles in speech-specific processing.

### **Hemisphere lateralization**

Speech processing in the current perspective is proposed to be bilateral but with differential sensitivity to specific spectro-temporal features of speech (Flinker et al., 2019; Giroud et al., 2020). Based on the asymmetric sampling in time (AST) hypothesis (Poeppel, 2003), high- and low-frequency neural oscillations may have different hemisphere lateralization. Specifically, high-frequency neural oscillations with a short temporal integration window are left-hemisphere dominant, while low-frequency neural oscillations with a long temporal integration window are right-hemisphere dominant (Morillon et al., 2012). Similarly, our results revealed left-hemispheric lateralization in high- $\gamma$  power envelope tracking and the right-hemispheric dominance in  $\theta$  phase decoding accuracy. Since our data were obtained directly from human auditory cortex using sEEG technology with a high temporal-spatial resolution, this study provided prevailing evidence to the AST theory.

The lateralization of speech processing at different frequencies may reflect a spectral processing hierarchy (Giroud et al., 2020). From this view, the fast high- $\gamma$  power matches the evoked onset response of the human auditory cortex (Elhilali et al., 2004), which may reflect a general response to the low-level acoustic features in speech. In contrast, the slow  $\theta$  phase may further select the specific linguistic properties of the speech. Combined, these two stages of processing may shed light on the neural dynamics in cortical auditory processing in the left and right hemispheres.

In sum, we revealed two stages of speech envelope tracking, which are differently modulated by speech intelligibility. The current findings suggest that our auditory system may use different pathways to process sounds with and without semantic information, which may be an important mechanism for human brain to resolve speech information under adverse auditory environments.

## MATERIALS AND METHODS

### *Participants*

Ten patients (6 males and 4 females) with the mean age of 30.9 years old (from 13 to 46 years old), who were undergoing neurosurgical treatment for refractory focal epilepsy in the Beijing Tiantan Hospital and the Sanbo Brain Hospital of Capital Medical University, participated in this study. They were implanted with intracranial stereo-electroencephalogram (sEEG) electrodes based on clinical evaluation for respective surgery. All patients were right-handed with a self-reported normal hearing and provided written informed consent to participate. All patients were right-handed native Mandarin Chinese speakers. Experimental procedures were approved by the Ethics Committee of the Beijing Tiantan Hospital and the Sanbo Brain Hospital of Capital Medical University.

### *Acoustic stimuli and experimental procedures*

The natural speech used in this study was a Chinese "nonsense" sentence recorded from a young female talker (Gao et al., 2017). The "nonsense" sentences are correct in syntactical level but semantically meaningless, which were usually used as stimuli for speech intelligibility measurement (Yang et al., 2007). In the current study, the English translation of the natural speech is "*This kind of Suzhou has remembered your butterfly*" (“这种苏州已经记住你的蝴蝶”).

The spectrally degraded speech stimulus was a reconstructed version of the natural speech using a custom-designed 2-band noise-vocoder in MATLAB (Shannon et al., 1995). In detail,

we first filtered the natural speech and a speech-spectrum noise (same power spectrum as the natural speech) into two bands (band 1: 80 to 1236 Hz; band 2: 1236 to 8820 Hz; spaced in equal steps along the cochlear frequency map) (Greenwood, 1990). Secondly, we extracted the envelope and the temporal fine structure (TFS) (Wang and Li, 2017) from the filtered signals of both the natural speech and a speech-spectrum noise. Thirdly, the envelope of each filter output from the natural speech was then multiplied by the TFS of the corresponding filter output from the speech-spectrum noise (Smith et al., 2002), resulting in two natural speech-noise chimeras. Finally, we summed these two chimeras (with energy normalized) to produce a noise-vocoded speech (NV). According to behavioral evidence (Newman and Chatterjee, 2013), noise-vocoded speech with 2 band is unintelligible. The correlation coefficient between the natural speech and the noise-vocoded speech was 0.99 (Figure 1).

----- Insert Figure 1 about here -----

Both the natural and the noise-vocoded speech were calibrated to the level of 70 dB SPL using a Larson Davis Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824). The duration of either stimulus was 3.275 s. The acoustic stimuli were transferred using Creative Sound Blaster (Creative SB X-Fi Surround 5.1 Pro, Creative Technology Ltd, Singapore) and presented to participants binaurally at a sampling rate of 22050 Hz with insert earphones (ER-3, Etymotic Research, Elk Grove Village, IL). To avoid the prior knowledge effect of the natural speech on the noise-vocoded speech, we always presented the noise-vocoded speech first. Both stimuli were repeated 30 times. The inter-stimulus interval (ISI) between the two stimulus presentations was random between 1800 to 2200 ms.

Patients were reclining on a bed in their private clinical suite during the experiment. No seizures were observed at least 4 hours before the experimental procedure in any patients. We asked the patients to stay awake and passively listen to the sounds without any behavioral task. sEEG data were recorded with a Nihon Kohden (Tokyo, Japan) clinical monitoring system using a sampling rate of 2000 Hz with the online bandpass filter of 0.08-600 Hz in Tiantan Hospital or using the Nicolet video-EEG monitoring system (Natus Neuro, USA) with the sampling rate of 512 Hz (bandpass filter of 0.05-200 Hz) in Sanbo Hospital.

### ***sEEG contact localization***

Each sEEG electrode (Huake Hengsheng Medical Technology Co Ltd, Beijing, China) had 8-16 macro contacts (0.8-mm diameter, 2.0-mm length, 1.5-mm apart from each other). All subjects underwent the pre-implant structural MRI (T1-weighted) and the post-implant CT scans. The T1 scan was firstly used to reconstruct the three-dimensional brain surfaces using BrainSuite (2018a) (<http://brainsuite.org/>), and then aligned to the post-implant CT using a rigid transformation in BioImage (<http://bioimagesuite.yale.edu>). We then identified the coordinates of sEEG electrodes using the Brainstorm toolbox (Tadel et al., 2011; <http://neuroimage.usc.edu/brainstorm/>) in the MATLAB environment. The anatomical of each contact was identified using an individual atlas (Joshi et al., 2020) of each patient. Contacts localized in the Heschl's gyrus (HG), superior temporal gyrus (STG) were selected. The sEEG electrodes have carved out the way to directly investigate the function of human auditory cortex in auditory perception (Fang and Hu, 2021). All the subject-wised coordinates were then transferred into MNI coordinates for group plotting.

### ***Electrophysiological data analyses***

**Preprocessing.** Raw sEEG data from each contact were imported into MATLAB (R2018a, MathWorks) and visually inspected for the presence of artifacts and interictal epileptic spikes (Lu et al., 2021), which were excluded from the further analyses. Raw sEEG data were pre-processed using the EEGLAB toolbox (Brunner et al., 2013) in the MATLAB environment. First, we extracted broadband signal (0.5-200 Hz) and high- $\gamma$  signal (60-140 Hz) from the recorded sEEG data. Next, broadband signals, were segmented into single-trial epochs from -1.3 s to 4.5 s around the sound onset. Meanwhile, the power envelope of the bandpass filtered high- $\gamma$  signals were extracted by Hilbert transform, segmented into single-trial epochs from -0.5 s to 4.5 s around the sound onset, down-sampled to 100 Hz and z-scored to the baseline. A contact was selected for the following analyses, if it had the significantly larger response power against baseline for a period lasting at least 10 ms (Bonferroni correction,  $p < 0.05$ ) for both the broadband signal and the high- $\gamma$  signal.

**High- $\gamma$  power envelope tracking.** The tracking strength between neural activity (high- $\gamma$  power) and the speech envelope was quantified by a cross-correlation function using the Matlab function *crosscorr*. Considering that the response lags the stimulus, we calculated the correlation coefficient between the averaged high- $\gamma$  power and the stimulus envelope (down-sampled to 100 Hz) were calculated for the different time lags (0 to 200 ms in step of 10 ms, the stimulus relative to the response). Confidence intervals were obtained by assuming high- $\gamma$  power and the stimulus envelope uncorrelated, if the maximum correlation coefficient at the optimal lag time was larger than the upper value of the confidence interval, it was identified as

statistically significant and kept for further comparisons. Otherwise, the correlation coefficient was replaced by the zero value.

**$\theta$  phase envelope tracking.** We quantified  $\theta$  phase envelope tracking using cerebro-acoustic coherence (CACoh) (Harding et al., 2019; Peelle et al., 2013), a measure of how sEEG signal phases align with speech envelope phases. Specifically, single-trial broadband signals were transformed to the time-frequency domain using the complex Morlet wavelets method with a frequency resolution of 0.5 Hz. The same time-frequency transformation was applied to the speech envelope. The phase coherence between single-trial broadband signal and the speech envelope was calculated for each of 14 frequency bins between 3 and 10 Hz according to the following formula:

$$CACoh_{(f)} = \frac{1}{t} \left| \sum_{k=1}^t e^{i\Delta\theta(f,k)} \right|$$

where  $\Delta\theta(f, k)$  is the angular distance between speech envelope and sEEG signal at a single time point ( $k$ ),  $e$  is the base of natural logarithms (~2.71),  $i$  is the imaginary unit that satisfies  $i^2 = -1$ , and  $t$  is time.

Since cortical tracking of speech envelope lags the acoustic signal (Di Liberto et al., 2015; Ding and Simon, 2012), we first calculated CACoh for time lags up to 400 ms (10-ms steps) between the speech envelopes and single-trial sEEG signal. Then, for each contact, we obtained the peak CACoh value at the optimal lag for each frequency bin and averaged them over trials, resulting in one CACoh per frequency bin for the natural speech and noise-vocoded speech. To test the statistical significance of CACoh results, we built a null distribution by randomly shifting the sEEG time series then calculating their CACoh 100 times. Actual CACoh values larger than 95% of the null distribution were considered significant, otherwise replaced by zero

in further comparisons. We then calculated the z-score transform of CACoh values based on the null distribution and averaged the results in  $\theta$  band (4-8 Hz).

**Spectra-temporal decoding.** To further investigate the relationship between the spectral contrast and neural oscillation activities, we used high- $\gamma$  band power and  $\theta$ -band phase to classify natural and noise-vocoded speech stimuli. We employed a support vector machine (SVM) approach using single trial time-frequency data from the high- $\gamma$  power or  $\theta$  phase, respectively. For high- $\gamma$  power decoding, the responses (0 to 4000 ms, relative to the stimulus beginning) were segmented into non-overlapping 50-ms time bins, and these 50-ms windows were averaged to extract power values served as the model feature for this type. For  $\theta$  phase decoding, responses were segmented into non-overlapping 100-ms time bins, and the model features were vectors of phase values by averaging across the 100-ms windows that were segmented. For each contact, we used the feature above separately to obtain model accuracy via a leave-one-out cross validation method. The SVM model was trained with a linear kernel and the penalty parameter set to 1. All analysis was performed in MATLAB (R2018b, MathWorks, MA, USA) using functions from the LIBSVM toolbox (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>). In order to test the significance of decoding values, we compared the observed decoding accuracy with a null distribution generated by permutation testing (swapping trial labels before SVM training). For each contact, the decoding accuracy with  $p$  value for high- $\gamma$  band and  $\theta$  band was obtained.

### *Statistical analyses*

Statistical analyses were performed with IBM SPSS Statistics 20 software (SPSS, Chicago, IL). Paired *t*-tests, independent-samples *t*-tests, chi-square test repeated-measures analysis of variance (ANOVAs) and post hoc comparisons (with Bonferroni corrections), and Pearson correlation tests were conducted. The null hypothesis rejection level was set at 0.05.

## Compliance and ethics

The author(s) declare that they have no conflict of interest.

## **Acknowledgements**

This study was supported by the National Key R&D Program of China (2017YFC1307500), National Natural Science Foundation of China (81771399, 32171039), Capital Healthy Development Research Funding (2016-1-2011, 2020-1-2013), and the Beijing Natural Science Foundation (Z200024).

## REFERENCES

Akbari, H., Khalighinejad, B., Herrero, J.L., Mehta, A.D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Sci Rep* 9, 874.

Brunner, C., Delorme, A., and Makeig, S. (2013). Eeglab - an Open Source Matlab Toolbox for Electrophysiological Research. *Biomed Tech (Berl)* 58 Suppl 1.

Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313, 1626-1628.

Clarke, S., and Morosan, P. (2012). Architecture, connectivity, and transmitter receptors of human auditory cortex. In: *The Human Auditory Cortex*, 2nd Edn., eds D. Poeppel, T. Overath, A. N. Popper, and R. R. Fay (New York, NY: Springer), pp. 11-38.

Davis, M.H., and Johnsrude, I.S. (2003). Hierarchical processing in spoken language comprehension. *J Neurosci* 23, 3423-3431.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen* 134, 222-241.

Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr Biol* 25, 2457-2465.

Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018). Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. *eNeuro* 5.

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

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

Doelling, K.B., and Poeppel, D. (2015). Cortical entrainment to music and its modulation by expertise. *Proc Natl Acad Sci U S A* 112, E6233-6242.

Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z., and Shamma, S.A. (2004). Dynamics of precise spike timing in primary auditory cortex. *J Neurosci* 24, 1159-1172.

Eliades, S.J., Crone, N.E., Anderson, W.S., Ramadoss, D., Lenz, F.A., and Boatman-Reich, D. (2014). Adaptation of high-gamma responses in human auditory association cortex. *J Neurophysiol* 112, 2147-2163.

Etard, O., and Reichenbach, T. (2019). Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. *J Neurosci* 39, 5750-5759.

Fang, F., and Hu, H. (2021). Recent progress on mechanisms of human cognition and brain disorders. *Sci China Life Sci* 64, 843-846.

Flinker, A., Doyle, W.K., Mehta, A.D., Devinsky, O., and Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nat Hum Behav* 3, 393-405.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., and Giraud, A.L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat Commun* 5, 4694.

Gao, Y., Wang, Q., Ding, Y., Wang, C., Li, H., Wu, X., Qu, T., and Li, L. (2017). Selective Attention Enhances Beta-Band Cortical Oscillation to Speech under "Cocktail-Party" Listening Conditions. *Front Hum Neurosci* 11, 34.

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., and Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. *PLoS Biol* 18, e3000207.

Gourevitch, B., Martin, C., Postal, O., and Eggermont, J.J. (2020). Oscillations in the auditory system and their possible role. *Neurosci Biobehav Rev* 113, 507-528.

Greenwood, D.D. (1990). A cochlear frequency-position function for several species--29 years later. *J Acoust Soc Am* 87, 2592-2605.

Harding, E.E., Sammler, D., Henry, M.J., Large, E.W., and Kotz, S.A. (2019). Cortical tracking of rhythm in music and speech. *Neuroimage* 185, 96-101.

Hauswald, A., Keitel, A., Chen, Y.P., Rösch, S., and Weisz, N. (2020). Degradation levels of continuous speech affect neural speech tracking and alpha power differently. *Eur J Neurosci*. (in press)

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

Howard, M.F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol* 104, 2500-2511.

Joshi, A.A., Choi, S., Chong, M., Sonkar, G., Gonzalez-Martinez, J., Nair, D., Wisniewski, J.L., Haldar, J.P., Shattuck, D.W., Damasio, H., et al. (2020). A Hybrid High-Resolution Anatomical MRI Atlas with Sub-parcellation of Cortical Gyri using Resting fMRI. *bioRxiv* 2020.09.12.294322.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PLoS One* 8, e53398.

Kulasingham, J.P., Brodbeck, C., Presacco, A., Kuchinsky, S.E., Anderson, S., and Simon, J.Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *Neuroimage* 222, 117291.

Lu, J., Luo, L., Wang, Q., Fang, F., and Chen, N. (2021). Cue-triggered activity replay in human early visual cortex. *Sci China Life Sci* 64, 144-151.

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001-1010.

Mai, G., Minett, J.W., and Wang, W.S. (2016). Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *Neuroimage* 133, 516-528.

Millman, R.E., Johnson, S.R., and Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J Cogn Neurosci* 27, 533-545.

Morillon, B., Liégeois-Chauvel, C., Arnal, L.H., Bénar, C.G., and Giraud, A.L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. *Front Psychol* 3, 248.

Newman, R., and Chatterjee, M. (2013). Toddlers' recognition of noise-vocoded speech. *J Acoust Soc Am* 133, 483-494.

Nourski, K.V., Steinschneider, M., Rhone, A.E., Kovach, C.K., Kawasaki, H., and Howard, M.A., 3rd. (2019). Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. *Hear Res* 371, 53-65.

Nourski, K.V., Steinschneider, M., McMurray, B., Kovach, C.K., Oya, H., Kawasaki, H., and Howard, M.A., 3rd. (2014). Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *Neuroimage* 101, 598-609.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., 3rd, and Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *J Neurosci* 29, 15564-15574.

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

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. *Speech Commun* 41, 245-255.

Poeppel, D., and Assaneo, M.F. (2020). Speech rhythms and their neural foundations. *Nat Rev Neurosci* 21, 322-334.

Rimmele, J.M., Zion Golumbic, E., Schröger, E., and Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex* 68, 144-154.

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

Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87-90.

Synigal, S.R., Teoh, E.S., and Lalor, E.C. (2020). Including Measures of High Gamma Power Can Improve the Decoding of Natural Speech From EEG. *Front Hum Neurosci* 14, 130.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., and Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *J Assoc Res Otolaryngol* 19, 181-191.

Wang, Q., and Li, L. (2017). Modelling envelope and temporal fine structure components of frequency-following responses in rat inferior colliculus. *Sci China Technol Sci* 60, 966-973.

Wang, Y., Zhang, L., Pan, J., Xie, K., Li, S., Wang, Z., and Lin, L. (2008). Ripple-associated high-firing interneurons in the hippocampal CA1 region. *Sci China Life Sci* 51, 120-126.

Yang, Z., Jing, C., Huang, Q., Wu, X., Wu, Y., Schneider, B.A., and Liang, L.J.S.C. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Commun* 49, 892-904.

Yellamsetty, A., and Bidelman, G.M. (2018). Low- and high-frequency cortical brain oscillations reflect dissociable mechanisms of concurrent speech segregation in noise. *Hear Res* 361, 92-102.

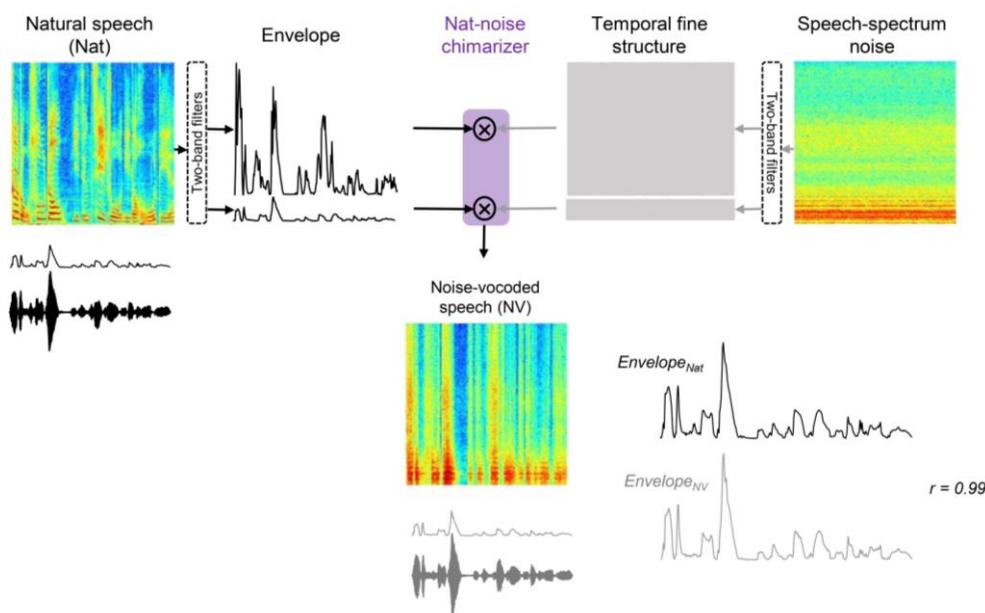
Yu, L.D., and Wang, S.P. (2021). Aberrant auditory system and its developmental implications for autism. *Sci China Life Sci* 64, 861-878.

Zhang, Y., Zhang, Y.Y., and Fang, F. (2020). Neural mechanisms of feature binding. *Sci China Life Sci* 63, 926-928.

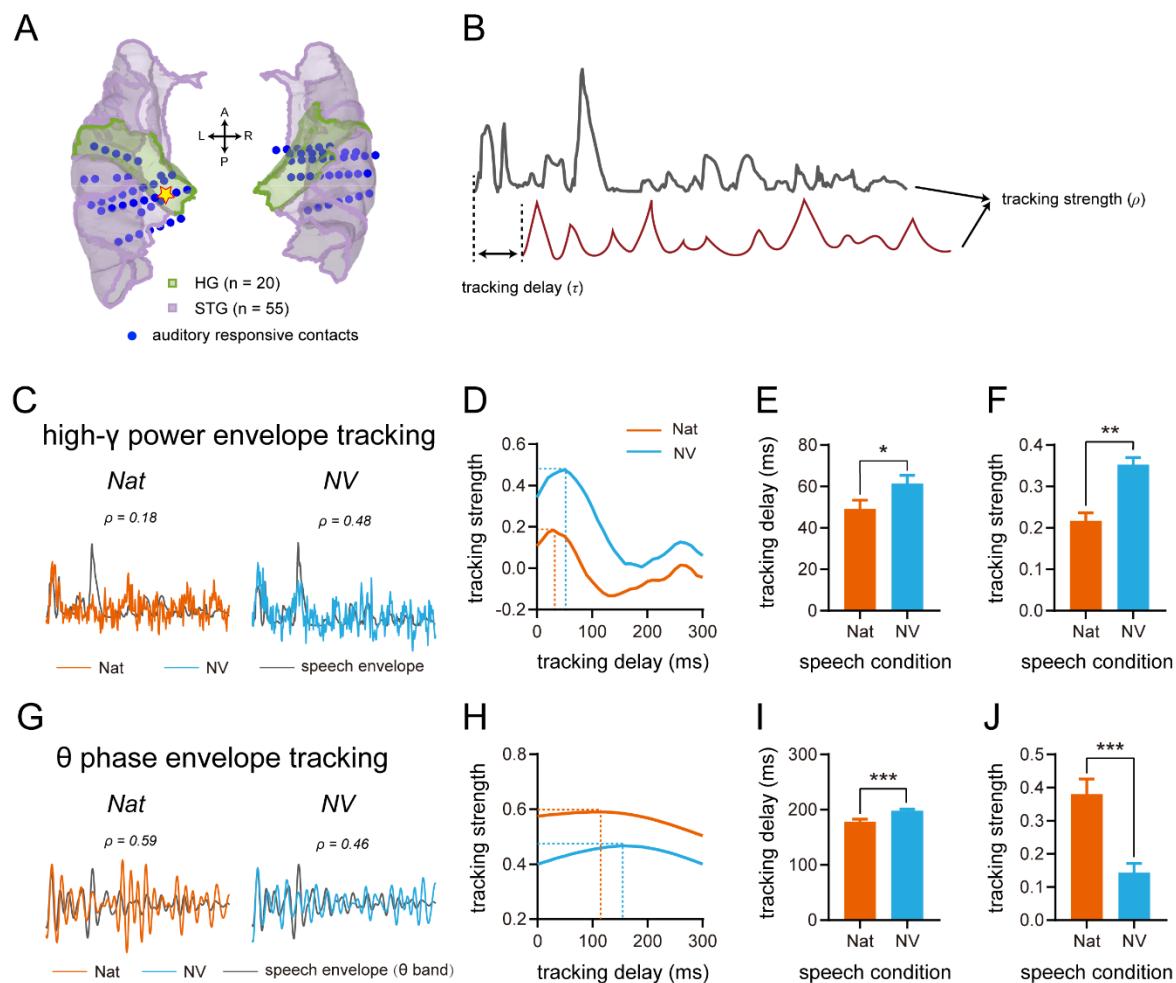
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron* 77, 980-991.

Zou, J., Xu, C., Luo, C., Jin, P., Gao, J., Li, J., Gao, J., Ding, N., and Luo, B. (2021). θ-Band Cortical Tracking of the Speech Envelope Shows the Linear Phase Property. *eNeuro* 8.

## FIGURE LEGENDS

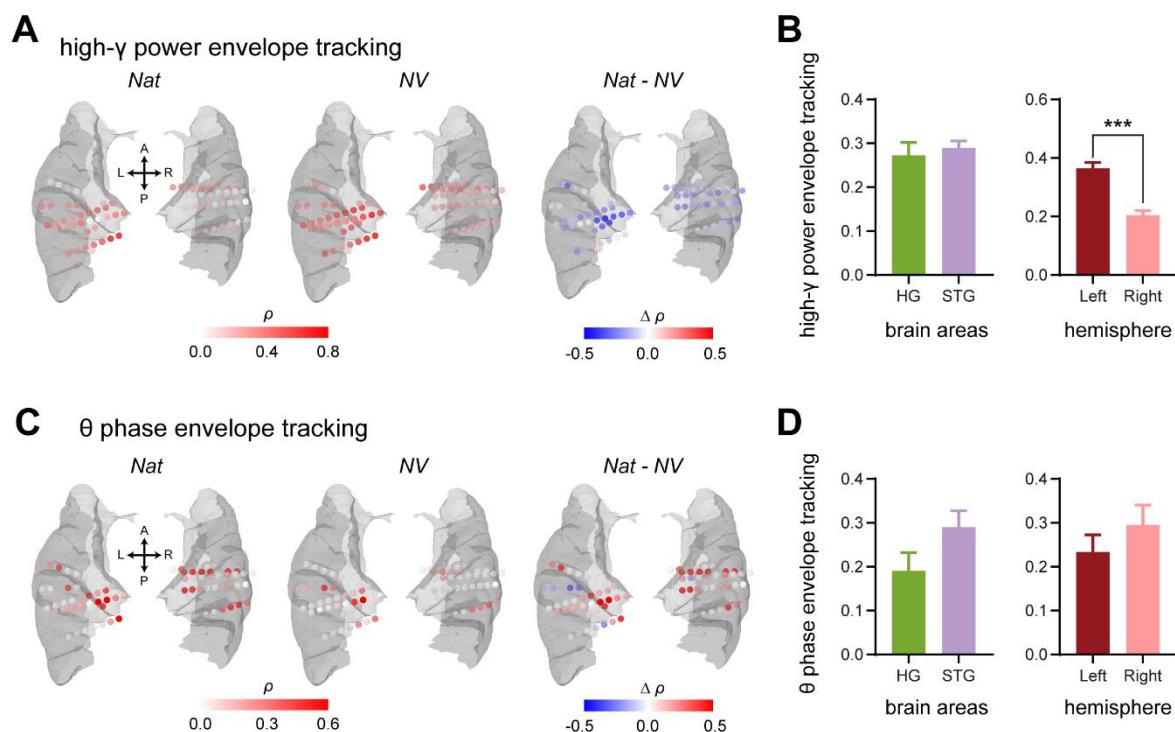


**Figure 1.** Schematic illustration of the noise-vocoded speech synthesis. Firstly, the natural speech and the speech-spectrum noise are filtered into two bands. Secondly, the outputs of these two bands are Hilbert transformed to extract the envelopes and the temporal fine structures (TFSs). Thirdly, the envelope of the natural speech is multiplied by the TFS of the speech-spectrum noise in each band. Finally, these two chimeras are summed to form the noise-vocoded speech (NV). Notably, the envelopes of the natural speech and that of the noise-vocoded speech are approximately identical (correlation coefficient = 0.99). *Nat, natural speech; NV, noise-vocoded speech.*

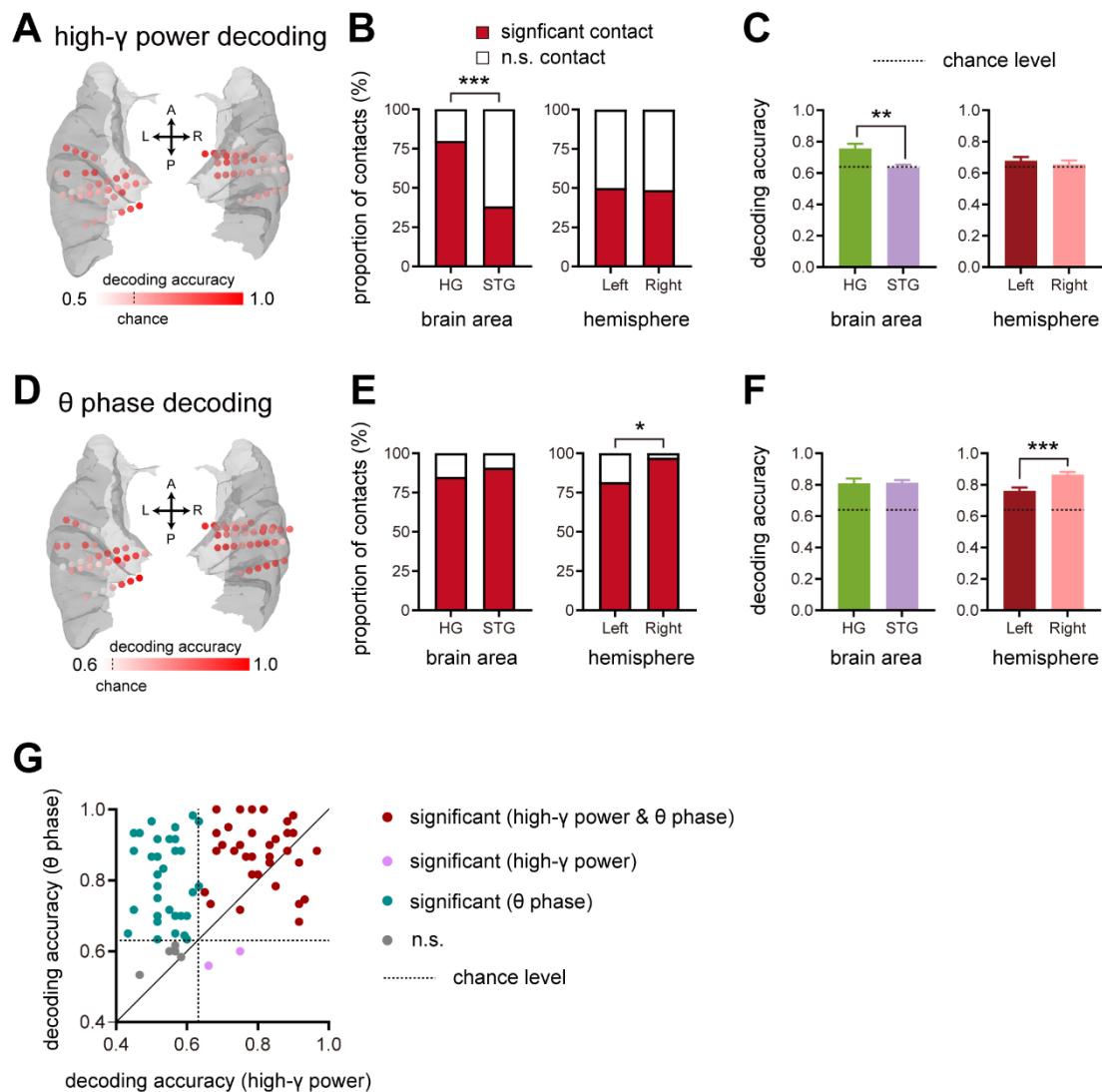


**Figure 2.** **A**, anatomical parcellation of temporal lobe regions of the human auditory cortex, and contacts selected from all patients counts across the selected anatomical areas and projected onto a Montreal Neurological Institute (MNI) atlas brain (BCI-DNI\_BrainSuite\_2016). HG, Heschl's gyrus (in green); STG, superior temporal gyrus (in purple), the exemplary contact is marked by a yellow five-pointed star with the red edge. **B**, the diagram showing the tracking strength ( $\rho$ ) of two example waveforms obtained at the optimal tracking delay ( $\tau$ ). **C**, the waveform of the Nat-speech envelope and the NV-speech envelope (gray curves), and the corresponding averaged high- $\gamma$  power responses (Nat, orange curves; NV, blue curves) of the example contact (five-pointed star in A). **D**, high- $\gamma$  power envelope tracking for Nat-speech and

NV-speech at different lag times are plotted for the example waveforms shown in **C**. The dashed lines show the maximum tracking strengths at the optimal delay times. **E**, comparison of the tracking delay for high- $\gamma$  power between the condition of Nat-speech and NV-speech. **F**, comparison of tracking strength for high- $\gamma$  power between Nat-speech and NV-speech. **G**, single-trial  $\theta$  band responses of the example contact (five-pointed star in **A**) for Nat-speech (orange curves) and NV-speech (blue curves), and the waveforms of the speech envelope (gray curves) filtered by  $\theta$  band (4-8 Hz). **H**,  $\theta$  phase envelope tracking for Nat-speech and NV-speech at different lag times are plotted for the example waveforms shown in **G**. **I**, tracking delay of  $\theta$  phase for Nat-speech and NV-speech is compared. **J**,  $\theta$  phase tracking strength is compared for Nat-speech and NV-speech. *Nat*, natural speech; *NV*, noise-vocoded speech.



**Figure 3.** Envelope tracking in high- $\gamma$  power and  $\theta$  phase. **A**, high- $\gamma$  power envelope tracking for Nat-speech (left subpanel), NV-speech (middle subpanel), and the difference between these two conditions (Nat-NV) (right subpanel) are plotted on the brain areas selected. **B**, Comparison of the high- $\gamma$  power envelope tracking for the speech stimuli (Nat- and NV-) between brain areas (left subpanel, HG vs. STG), and between left and right hemispheres (right subpanel). **C**,  $\theta$  phase envelope tracking for Nat-speech (left subpanel), NV-speech (middle subpanel), and the difference between these two conditions (Nat-NV) (right subpanel) are plotted on the brain. **D**,  $\theta$  phase envelope tracking is compared for the speech stimuli (Nat- and NV-) between brain areas (left subpanel, HG vs. STG) and between left and right hemispheres (right subpanel). *Nat, natural speech; NV, noise-vocoded speech.* \*\*\* $p < 0.001$ .



**Figure 4.** Decoding of Nat-speech and NV-speech using the feature of either high- $\gamma$  power or  $\theta$  phase. **A**, distributions of high- $\gamma$  power decoding accuracy for each contact in the selected brain areas. **B**, proportions of the significant and the non-significant (n.s.) contacts for high- $\gamma$  power decoding accuracy across brain areas and hemispheres, respectively. **C**, comparisons of high- $\gamma$  power decoding accuracy between brain areas (left subpanel, HG vs. STG) and between left and right hemispheres (right subpanel), respectively. **D**, distributions of  $\theta$  phase decoding accuracy for each contact in the selected brain areas. **E**, proportions of the significant and the

non-significant (n.s.) contacts for  $\theta$  phase decoding accuracy across brain areas and hemispheres, respectively. **F**,  $\theta$  phase decoding accuracy between brain areas (left subpanel, HG vs. STG) and between left and right hemispheres (right subpanel) are compared respectively. **G**, scatter plot shows decoding accuracy using high- $\gamma$  power feature against the decoding accuracy using  $\theta$  phase feature for each contact. Values with different colors labeled show the statistical significance of their decoding accuracy. Dotted lines in this figure denote the chance level obtained by permutation testing ( $n = 1000$ ). Decoding values above these dotted lines are higher than 95% of random values. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ .