Received: 28 April 2023 | Revised: 17 August 2023 | Accepted: 13 September 2023

DOI: 10.1002/hbm.26503

RESEARCH ARTICLE

WILEY

Distinct roles of delta- and theta-band neural tracking for sharpening and predictive coding of multi-level speech features during spoken language processing

Guangting Mai[1,2,3] | William S.-Y. Wang[4,5]

[1]Hearing Theme, National Institute for Health Research Nottingham Biomedical Research Centre, Nottingham, UK

[2]Academic Unit of Mental Health and Clinical Neurosciences, School of Medicine, The University of Nottingham, Nottingham, UK

[3]Division of Psychology and Language Sciences, Faculty of Brain Sciences, University College London, London, UK

[4]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hung Hom, Hong Kong

[5]Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong, China

Correspondence
Guangting Mai, Hearing Theme, National Institute for Health Research Nottingham Biomedical Research Centre, Ropewalk House, 113 Ropewalk, Nottingham, UK.
Email: guangting.mai@nottingham.ac.uk

Abstract

The brain tracks and encodes multi-level speech features during spoken language processing. It is evident that this speech tracking is dominant at low frequencies (<8 Hz) including delta and theta bands. Recent research has demonstrated distinctions between delta- and theta-band tracking but has not elucidated how they differentially encode speech across linguistic levels. Here, we hypothesised that delta-band tracking encodes prediction errors (enhanced processing of unexpected features) while theta-band tracking encodes neural sharpening (enhanced processing of expected features) when people perceive speech with different linguistic contents. EEG responses were recorded when normal-hearing participants attended to continuous auditory stimuli that contained different phonological/morphological and semantic contents: (1) real-words, (2) pseudo-words and (3) time-reversed speech. We employed multivariate temporal response functions to measure EEG reconstruction accuracies in response to acoustic (spectrogram), phonetic and phonemic features with the partialling procedure that singles out unique contributions of individual features. We found higher delta-band accuracies for pseudo-words than real-words and time-reversed speech, especially during encoding of phonetic features. Notably, individual time-lag analyses showed that significantly higher accuracies for pseudo-words than real-words started at early processing stages for phonetic encoding (<100 ms post-feature) and later stages for acoustic and phonemic encoding (>200 and 400 ms post-feature, respectively). Theta-band accuracies, on the other hand, were higher when stimuli had richer linguistic content (real-words > pseudo-words > time-reversed speech). Such effects also started at early stages (<100 ms post-feature) during encoding of all individual features or when all features were combined. We argue these results indicate that delta-band tracking may play a role in predictive coding leading to greater tracking of pseudo-words due to the presence of unexpected/unpredicted semantic information, while theta-band tracking encodes sharpened signals caused by more expected phonological/morphological and semantic contents. Early presence of these effects reflects rapid computations of sharpening and prediction errors. Moreover, by measuring changes in EEG alpha

power, we did not find evidence that the observed effects can be solitarily explained by attentional demands or listening efforts. Finally, we used directed information analyses to illustrate feedforward and feedback information transfers between prediction errors and sharpening across linguistic levels, showcasing how our results fit with the hierarchical Predictive Coding framework. Together, we suggest the distinct roles of delta and theta neural tracking for sharpening and predictive coding of multi-level speech features during spoken language processing.

KEYWORDS
delta and theta bands, neural sharpening, neural tracking of speech, phonological/morphological processing, predictive coding, semantic processing

1 | INTRODUCTION

The brain neurally tracks and encodes various speech features as we perceive spoken languages. The neural tracking of speech describes the alignment between brain signals and important speech features which is associated with improvements in speech comprehension (Wöstmann, Fiedler, & Obleser, 2017). Speech tracking is dominated by low-frequency (<8 Hz) neural oscillations which encode features at different hierarchical levels, from acoustics to phonology and semantics (Di Liberto et al., 2015), enabling us to recognise words and understand speech (Lesenfants et al., 2019). The neural tracking not only depends on sensory inputs of speech but is also modulated by processing of higher-level linguistic content (Broderick et al., 2019; Coopmans et al., 2022; Donhauser & Baillet, 2020; Gross et al., 2013; Kaufeld et al., 2020; Keitel et al., 2018; Mai et al., 2016) as well as participants' language experience (Etard & Reichenbach, 2019; Tezcan et al., 2023). For example, greater low-frequency neural tracking is related to existence of phonological information (normal vs. time-reversed speech, Gross et al., 2013; Mai et al., 2016), greater semantic similarity (Broderick et al., 2019), richer lexical content (Coopmans et al., 2022) and compositional structure (coherent lexical-syntactic information, Kaufeld et al., 2020). It is also shown that low-frequency neural tracking of speech is modulated by language experience, for instance, greater neural tracking of higher-level (phoneme), rather than low-level (acoustic-edge), features were observed when participants perceived native compared to when they perceived an unfamiliar non-native language (Tezcan et al., 2023).

During these higher-level modulatory processes, prior expectation is an important factor that further affects speech tracking. For example, more expected speech content (e.g., higher speech intelligibility, greater semantic expectedness, or greater acoustic clarity) enhances low-frequency neural tracking of speech (Broderick et al., 2019; Coopmans et al., 2022; Etard & Reichenbach, 2019; Peelle et al., 2013; Tezcan et al., 2023). On the other hand, there is also evidence showing greater neural tracking of stimuli with greater unexpectedness or surprisal (Donhauser & Baillet, 2020; Sohoglu & Davis, 2020). The current study focuses on the effects of prior expectation during speech perception and considers how expected and unexpected linguistic content may modulate speech tracking. Previous research has demonstrated two mechanisms for the processing of linguistic content with different degrees of expectation for neural tracking of speech. The first states that neural representations for input stimuli are enhanced or 'sharpened' by prior expectations embedded in higher-order cognitive or linguistic processes (de Lange et al., 2018). This would mean greater tracking when speech contents are more expected linguistically (e.g., Broderick et al., 2019). The second states that expected stimulus inputs are minimally encoded and subtracted so that more unexpected information is represented as prediction errors (Arnal & Giraud, 2012; Blank & Davis, 2016; Friston, 2005, 2012; Sohoglu & Davis, 2016; Summerfield & De Lange, 2014). This would mean greater tracking when more unexpected contents are present within speech signals (Donhauser & Baillet, 2020; Sohoglu & Davis, 2020). It is suggested that the two mechanisms coexist and can be detected in the same neural responses to speech (Broderick & Lalor, 2020). Here, we ask a research question of whether electrophysiological responses to continuous speech can reflect neural sharpening and predictive coding respectively at different frequency ranges, particularly delta and theta bands.

Indeed, previous research has shown differences between delta- and theta-band tracking of speech. For example, Ding et al. (2014) showed that delta-band tracking can be enhanced in more difficult listening situations (listening to noise-vocoded speech with lower spectral resolutions) possibly reflecting greater listening efforts, while the reversed pattern was found for theta-band tracking. Etard and Reichenbach (2019) manipulated the speech comprehensibility (native vs. non-native language) and clarity (presented in quiet vs. in noise) and showed that greater delta-band tracking contributes to better comprehension while greater theta-band tracking reflects better encoding of speech clarity. Donhauser and Baillet (2020) showed that both delta- and theta-band tracking are involved during predictive coding with the former more related to phonemic surprisal (unexpectedness of the presence of a phoneme) and the latter more related to contextual uncertainty (uncertainty for the upcoming phonemes). While it is indicative that the roles of delta- and theta-band tracking are distinct, it is uncertain whether their roles are different for neural sharpening and predictive coding, for example, whether one plays greater roles for sharpening/predictive coding compared to the other,

and whether these different roles (if there are) can be observed for both sharpening and predictive coding in the same neural dataset.

Our previous study (Mai et al., 2016) has shown the differences in delta- and theta-band tracking of speech with different degrees of expected linguistic content. In this study (Mai et al., 2016), participants listened to continuously spoken languages with stimuli containing different phonological and semantic contents (real-words, pseudo-words and time-reversed speech). Compared to time-reversed speech, both real-words and pseudo-words contained valid phonological/morphological contents. Real-words included semantically valid words while pseudo-words included invalid words (hence unexpected semantic information). We showed that phonological/morphological contents (real-words and pseudo-words vs. time-reversed speech) enhanced the theta-band tracking of speech envelopes, while unexpected semantic information (pseudo-words vs. real-words) enhanced delta-band tracking of envelopes. This indicates the potential distinctive roles of neural speech tracking between delta and theta bands for sharpening and predictive coding. However, it is still unclear (1) whether such distinctions are represented by tracking of features across multiple levels beyond speech envelopes (e.g., phonetics and phonemes); (2) when such distinctions occur across the processing stages after feature onsets; and (3) whether the greater tracking of stimuli with unexpected semantic information can actually be explained by greater listening efforts in difficult listening situations (Ding et al., 2014) instead of the predictive coding proposal.

In this current study, we used multivariate Temporal Response Functions (mTRF) (Crosse et al., 2016; Di Liberto et al., 2015) to analyse electroencephalographic (EEG) data when participants listened to speech with different linguistic contents in Mai et al. (2016) and measure encoding accuracies as representations for neural tracking of multi-level speech features. Neural sharpening scheme anticipates that richer linguistic content (hence with greater expectation) should result in higher encoding accuracies (i.e., real-words > pseudo-words > time-reversed speech, Figure 1a). Predictive coding, on the other hand, suggests that discrepancies between the heard stimuli and expected/predicted signals (i.e., prediction errors) are encoded (Sohoglu et al., 2012; Sohoglu & Davis, 2016, 2020) so that utterances containing unexpected information (i.e., pseudo-words) should yield the highest accuracy (Figure 1a). Note that we anticipate that time-reversed speech would always yield the lowest accuracy regardless of sharpening or predictive coding (or as low as real-words for predictive coding), since previous studies showed that neural tracking of forward speech is significantly greater than time-reversed speech at both delta and theta ranges (Gross et al., 2013; Mai et al., 2016). This would mean neural sharpening by phonological/morphological contents (forward vs. time-reversed speech) would coexist with prediction errors even when the predictive coding proposal is upheld. We here hypothesised that neural tracking of speech plays distinct roles at delta and theta bands for sharpening and predictive coding, where delta-band tracking plays roles in predictive coding driven by higher-level semantic processing while theta-band tracking plays roles in neural sharpening driven by richer phonological/morphological and semantic contents (Figure 1a). Beyond tracking of acoustic envelopes, we looked into encoding of more complex acoustic (spectrogram), phonetic and phonemic features. We first sought to examine whether the results can be explained by attention/listening effort. Due to the active sound-matching task of this study (see Section 2.2), we anticipate that stimuli with less rich linguistic information would grab more attention hence greater listening effort (e.g., Reichenbach et al., 2016) that results in greater neural tracking (i.e., real-words < pseudo-words

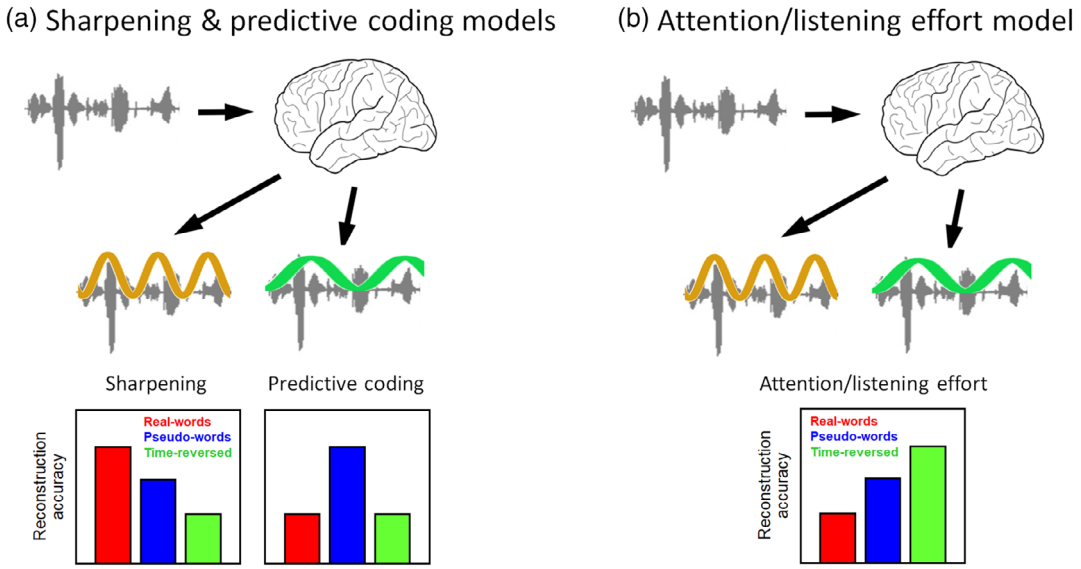

FIGURE 1 Predictions of results for the neural sharpening and predictive coding proposals (left) and the attention/listening effort proposal (right). The neural sharpening proposal anticipates that neural tracking should be greater for speech with richer (i.e., more expected) linguistic content (real-words > pseudo-words > time-reversed speech). The predictive coding proposal anticipates that pseudo-words which contain unexpected semantic information should yield the highest neural encoding accuracies. The attention/listening effort proposal anticipates speech with fewer linguistic content should grab more attention and lead to greater neural tracking (real-words < pseudo-words < time-reversed speech).

< time-reversed speech, Figure 1b). We then examined neural encoding accuracies across individual time lags to see what processing stages after feature onsets at which sharpening and prediction errors (or effects of attention/listening effort if attention models can explain the results) start to appear. Finally, we conducted directed information transfer analyses to illustrate how our findings may fit with the hierarchical Predictive Coding framework (Friston, 2005, 2012).

2 | MATERIALS AND METHODS

2.1 | Participants

Twenty normal-hearing (audiometric thresholds ≤25 dB HL across 0.5–6 kHz), native Mandarin speakers (8 males; aged 19–25 years) were recruited to participate in the experiment. No history of hearing, neurological or language disorders was reported for any participant. They were all right-handed (18 participants with handedness indices (HI) > 40 classified as strong right-handed and two participants with HIs = 33.3 classified as ambidextrous but towards right-handed) according to the Edinburgh Handedness Inventory (Oldfield, 1971). All participants were recruited and paid for with formal consents under approval of the Research Ethics Committee of The Chinese University of Hong Kong.

2.2 | Stimuli and tasks

Stimuli consisted of three types of continuous Mandarin utterances: (1) real-words, (2) pseudo-words, and (3) time-reversed speech. (1) and (2) were naturally produced by a male native Mandarin speaker recorded at a sampling rate of 22,050 Hz. All were produced with a syllable rate at ~4 Hz with all syllables having a similar duration of ~250 ms (except for the particle '的' which is ~150–200 ms, Figure 2a).[1] Each real-word utterance consisted of four semantically valid words with a syntactic structure of 'Subject + Verb + Attribute + [particle] + Object'. The words within an utterance were not contextually related to each other to keep participants' attention to the entire utterance during the behavioural tasks (see the next two paragraphs). A sample (translated) utterance is 'knowledge purchases fundamental opportunity', in which each word is a Mandarin disyllabic word (see Figure 2a). Pseudo-words were utterances consisting of the same number of morphologically valid syllables as in each real-word utterance, but with no two adjacent syllables forming a semantically valid word (Figure 2a). All participants confirmed after the experiment that all pseudo-word utterances were not semantically valid for them. It is important to note that the current pseudo-words are different from the commonly used 'Jabberwocky' stimuli where constituent syllables within the pseudo-words often do not have explicit meanings and/or potentials to form a valid word with another syllable (Kaufeld et al., 2020; Matchin et al., 2017, 2019; Pallier et al., 2011). In Mandarin (the language we used here), almost *all* morphologically valid syllables are commonly used morphemes that by themselves have some semantic meanings and each has potentials to form a valid disyllabic word with another morpheme.[2] This thus ensures word-level semantic priming, that is, the first syllable within each disyllabic word should provide *prior knowledge* to predict the second syllable, so that we could examine how neural tracking was affected by prediction/expectation. Time-reversed speech was temporally reversed versions of the real-words and pseudo-words. Time reversal causes substantial phonological distortion but retains similar acoustic complexity (temporal fluctuations, formant distributions and harmonic structures) as well as phonetic features of the original speech (Binder et al., 2000; Gross et al., 2013; Londei et al., 2010; Saur et al., 2010). There were 80 different utterances for each stimulus type (i.e., 240 sentences in total). Half of the time-reversed utterances were generated from real-words with the other half from pseudo-words. All stimuli had a similar duration (2.2–2.3 s) and were adjusted to the same average sound (root-mean-squared) intensity.

The experiment followed a within-subject design for which each participant was exposed to all three stimulus types. During the experiments, participants were seated in front of a computer screen and listened to the stimuli via EARTONE 3A inserted earphones (Etymotic Research, USA) with a fixed adjusted intensity at ~70 dB SPL for all utterances. All stimuli utterances were presented using EPrime 2.0 (Psychology Software Tools) and were divided into 8 blocks (i.e., 30 utterances with 10 for each stimulus type) with breaks taken in-between. The utterances in each block were presented in a randomised order. An additional practise block (30 utterances with different contents from the formal test) were run prior to the formal test.

The paradigm of each trial is shown in Figure 2b. At the start of each trial, there was a 3-s silence allowing participants to blink, followed by another 1.5-s silence with a white cross centred on the screen. A cue sound (200–300 ms long; a naturally produced syllable for the real-words and pseudo-words, or a time-reversed syllable for time-reversed speech) was then presented. These were followed by a 2-s silence and then the target utterance. Participants were required to complete a sound-matching task, in which they made a forced-choice judgement whether the cue sound was present in the target utterance or not by pressing a button after the end of the utterance (instructed by a question mark on the screen). They were instructed to sit still, keep their eyes on the white cross and avoid eye blinking or body movements as much as possible after the cue sound was played. They were also asked to press the button *only* after the question mark appeared to avoid motor artefacts during the target period. Feedback of accuracies was given after each block and participants were encouraged to respond as accurately as possible. Overall, the aim of the sound-matching task was to keep participants actively attending to the target utterances.

[1]Despite similar duration for all syllables, the utterances largely reserved the speech naturalness (from participant feedback post experiment) thanks to the syllable-timed nature of Mandarin (Mok, 2009).

[2]An example is the disyllabic word '基本' ('*fundamental*') embedded in the current example real-word utterance, which consists of syllables '基' (meaning 'base') and '本' (meaning 'origin/root'), while a corresponding pseudo-word could be '基米' in which '米' means 'rice' or 'metre'.

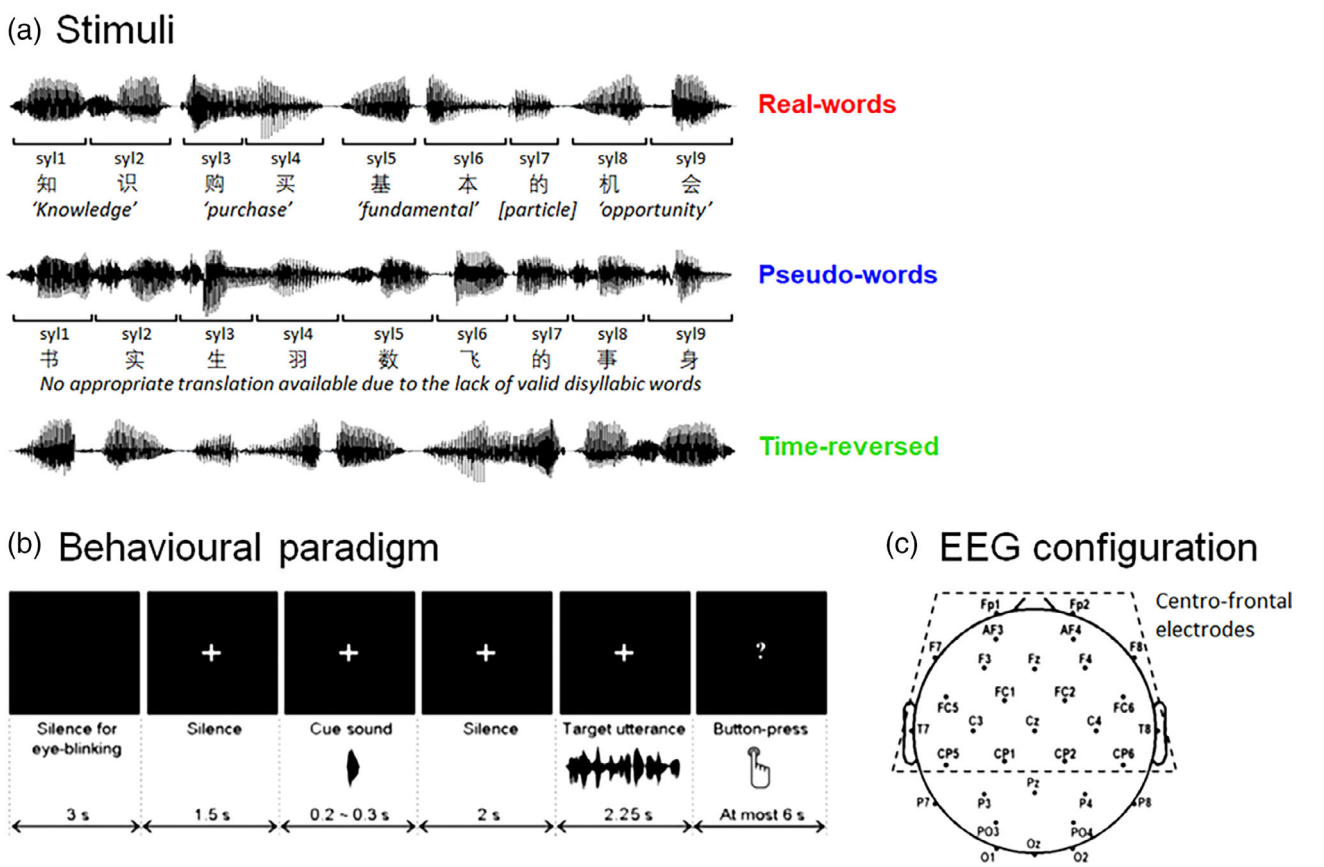

FIGURE 2 Experiment design. (a) Stimuli. Each trial of forward speech (real-words and pseudo-words) contained nine morphologically valid syllables (the seventh syllable is always the particle '的' that syntactically connects an adjective with a noun). Real-words are sentences consisting of four semantically valid disyllabic words which are not contextually related but fit with a fixed syntactic structure 'Subject + Verb + Attribute + 的 + Object' (a translated example '*Knowledge purchases fundemental opportunity*'). Pseudo-words are utterances with no adjacent syllables form a valid disyllabic word. Time-reversed speech are time-reversals of the forward speech. (b) Experiment paradigm. Participants were presented with a cue sound before the target speech and were required to press a button to judge whether the cue sound was present in the target speech. (c) EEG configuration. We focused on EEG reconstruction accuracies at centro-frontal electrodes (encompassed by the dashed trapezoid) covering temporal, frontal and parietal regions that are most important for auditory speech perception.

Out of all 80 utterances in each stimulus type, 20% of them in which the cue sounds were actually present in the target utterances (i.e., 16 utterances). In the present study, only the trials where the cue sounds were *not* present in the target utterances (i.e., 64 utterances) were included in the subsequent analyses. This was to preclude the possibility of participants not attending to the entire utterance period and to avoid the effects of target detection (e.g., P300, see van Dinteren et al., 2014) when the cue sound (as the target to be detected) was present in the utterance. This could also minimise possible effects of motor preparation of button press due to judgements made before the end of the utterance.

2.3 | EEG acquisition

2.3.1 | Acquisition and pre-processing

Scalp EEGs were recorded by a 32-electrode ActiveTwo system (Biosemi, The Netherlands) with layout consistent with the standard 10–20 system (see Figure 2c) and were sampled at 1024 Hz. CMS and DRL were used as ground electrodes. Bilateral mastoids were used as the reference. Eye artefacts were detected via vertical (vEOG; electrodes above and below the left eye) and horizontal EOGs (hEOG; electrodes on the lateral sides of the left and right eyes). Electrode offsets were always kept below 40 mV for all electrodes to ensure good quality of electrode contacts.

EEGs were pre-processed using Matlab R2022a (Mathworks). Signals of all electrodes (including EOGs) were first re-referenced to the bilateral mastoids and then bandpass filtered at 0.7–8 Hz using a zero-phase, second-order Butterworth filter. Signals for detecting eye artefacts were then obtained by subtracting between signals in corresponding EOG electrodes (vEOGs and hEOGs for vertical and horizontal artefacts, respectively). Trials, where the filtered EEGs in the target period (target utterances with a fixed length of 2.25 s for all trials) exceeded ±35 μV in any electrode (including vEOG and hEOG), were treated as being contaminated by eye or body movement artefacts and were rejected from subsequent analyses. Ultimately, out of the total 64 trials for each stimulus type (see Section 2.2), 60.75 ± 0.76

(mean ± standard error across participants), 59.95 ± 0.71 and 60.35 ± 0.85 trials were retained for real-words, pseudo-words and time-reversed speech, respectively. This means the rejection rates were 5.1 ± 1.2%, 6.3 ± 1.1% and 5.7 ± 1.3%, respectively.

Besides the approach of rejecting trials with artefacts, artefact corrections through Independent Component Analyses (ICA) were also conducted to correct vertical and horizontal eye blinks and other spurious components (e.g., components with large magnitudes appearing at a single electrode). However, as we found that EEG reconstruction accuracies after ICA corrections were, on average, lower than those after following the approach of trial rejections. This probably means that artefact correction has not resulted in the level of signal quality as in the artefact-free trials. Due to the relatively low rejection rate (<7% on average for all stimulus types), we decided to use the pre-processed signals based on trial rejections.

2.3.2 | Delta- and theta-band EEGs

Neural tracking was computed via multivariate Temporal Response Functions (mTRF; Di Liberto et al., 2015; Crosse et al., 2016; see Section 2.5 for details) by linearly mapping stimulus features onto EEG responses at the delta and theta range. The algorithm was applied for delta- and theta-band tracking separately for the three stimulus types in each participant.

To obtain neural signals at the delta and theta range, pre-processed EEGs were initially bandpass filtered into three frequency ranges: (1) 0.75–1.5 Hz ('ultra-low' delta), (2) 1.5–3 Hz (delta) and (3) 3–6 Hz (theta) using a zero-phase, second-order Butterworth filter. These particular numbers were chosen so that the mean cycles of the upper and lower bounds (667 ms and 1.333 s for ultra-low delta, 333 and 667 ms for delta and 167 and 333 ms for theta) correspond to the typical delta and theta cycles at 1 Hz (1 s per cycle), 2 Hz (500 ms per cycle) and 4 Hz (250 ms per cycle), respectively. The signals were then decimated to 64 Hz via a 30th-order Hamming-windowed FIR filter. The delta and theta EEGs for the artefact-free trials were then used for quantifying delta- and theta-band tracking, respectively. Subsequent analyses did not find above-chance/null tracking (tracking based on shuffled trial correspondence between EEG and speech stimuli) of targeted speech features (spectrogram plus phonetic and phonemic features) in the ultra-low delta-band for any stimulus type. We hence focussed on the delta and theta range in the present study.

2.4 | Extraction of stimulus features

The following stimulus features were extracted for each utterance: acoustic features of (1) spectrogram, (2) derivatives of spectrogram and (3) spectrotemporal modulations; higher-level phonetic and phonemic features. The length of the features was fixed at 2.25 s for all trials (as mentioned, all trials had durations at 2.2–2.3 s).

2.4.1 | Acoustic features

Acoustic features were extracted using Matlab 2022a. The spectrogram was obtained by first filtering each stimulus into 26 frequency channels between 100 and 5000 Hz using Gammachirp auditory filters which simulate the auditory periphery using the open-accessed Auditory Modelling Toolbox (Majdak et al., 2022; https://www.amtoolbox.org/). Bandwidth of each frequency band corresponded to one Equivalent Rectangular Bandwidth (ERB; Glasberg & Moore, 1990). Filtered waveforms in each channel were then Hilbert transformed to obtain the envelopes followed by further low-pass filtering at 30 Hz using a zero-phase, second-order Butterworth filter. In addition, we also bandpass filtered the envelopes in each channel into the corresponding frequency ranges of EEGs: (i) delta (1.5–3 Hz) and (ii) theta (3–6 Hz), so that EEGs can be reconstructed by spectrograms fluctuating at the corresponding frequency ranges (i.e., delta-band EEG reconstructed by delta envelopes while theta-band EEG reconstructed by theta envelopes). The derivatives of spectrogram were the first derivatives of the low-passed spectrogram followed by half-wave rectification. These reflect the acoustic onsets in the spectrogram, which have been shown to significantly contribute to EEG tracking (Daube et al., 2019).

Our subsequent analyses showed that both delta- and theta-band EEGs are best reconstructed by the spectrogram (with envelopes low-passed at 30 Hz; (see Section 3.2.1). Also, we found that combining spectrogram and the derivatives had not led to numerically higher reconstruction accuracies than spectrogram alone at the group level. We therefore used the spectrogram as the acoustic representation for EEG reconstruction.

In addition, Sohoglu and Davis (2020) also showed that neural encoding of spectrotemporal modulations (Elliott & Theunissen, 2009) could best model cortical responses. Other neurophysiological and fMRI studies have shown that cortical responses can be well-modelled using spectrotemporal modulations (Daube et al., 2019; Pasley et al., 2012; Santoro et al., 2014). We therefore further extracted such features following Sohoglu and Davis (2020). Specifically, the NSL toolbox in Matlab (Chi et al., 2005; http://nsl.isr.umd.edu/downloads.html) computed 128-channel auditory spectrogram (logarithmic-centred frequencies from 180 to 7040 Hz) for each utterance which was then wavelet filtered to extract spectral modulations of 0.5, 1, 2, 4 and 8 cycles per octave and temporal modulations of 1, 2, 4, 8 and 16 Hz. Other parameters were as follows: both frame length and time constant set at 15.6 ms to coordinate with the minimal interval of any two sampling points of decimated EEG (with 64 Hz sampling rate) and no linear compression. Modulations were finally averaged across the frequency channels and positive/negative temporal modulation directions to yield 25 features (5 frequency × 5 temporal modulations; Sohoglu & Davis, 2020). The features were used to model EEG via mTRF. However, we subsequently found that the EEG reconstruction accuracies using these features were not significantly above chance (null/shuffled accuracies) for any stimulus type for either delta- or theta-band tracking. This means spectrotemporal modulations would

not be appropriate acoustic representations for our current dataset. We thus focused on the spectrogram as the acoustic representation.

2.4.2 | Phonetic and phonemic features

To extract phonetic and phonemic features, onsets and offsets of phonemes were first annotated for each utterance for real-words and pseudo-words. Unlike previous studies which applied a forced-alignment algorithm (e.g., Di Liberto et al., 2015), there is no such tool for Mandarin. We therefore completed the annotation manually using the TextGrid function in Praat instead (University of Amsterdam, The Netherlands; see https://www.fon.hum.uva.nl/praat/manual/Intro_7__Annotation.html). The phonemes include 24 consonants and 23 vowels (7 single vowels, 12 diphthongs and 4 triphthongs) based on Standard Chinese phonology (Duanmu, 2007). Diphthongs and triphthongs were used as separate phonemic categories rather than consecutive single vowels because it is very difficult to define vowel boundaries due to their highly dynamic nature in continuous speech (most diphthongs and triphthongs are comprised of formant transitions without any acoustically steady sub-segments). Each phoneme vector was set at ones over the corresponding periods of that phoneme, otherwise at zeros.

We then converted the phonemic features into 19 phonetic features. These included 5 places (Labial, Dental, Alveolar, Retroflex, Velar) and 7 manners of articulations (Unaspirated Plosive, Aspirated Plosive, Unaspirated Affricate, Aspirated Affricate, Fricative, Nasal, Liquid) for consonants, 3 backnesses (Front, Central, Back), 3 heights (High, Medial and Low) and 1 mouth roundedness (Round) for vowels. Each phonetic feature vector was set at ones over the periods of corresponding phonemes (except for diphthongs and triphthongs which were weighted averages across component vowels), otherwise zeros. The weights of component vowels in diphthongs and triphthongs were based on the components' relative estimated durations. Specifically, for diphthongs /ai/, /aʊ/, /əʊ/ and /ei/ (in *IPA*), the weights were set the same for both component vowels (i.e., at 1/2). Other diphthongs all have glides of /i/, /u/ or /y/ preceding the nuclei (e.g., glide /i/ preceding nucleus /a/ in /ia/). These glides have relatively shorter durations than the nuclei (Duanmu, 2007) and we set the weights at one-third for the glides and two-thirds for the nuclei. For the triphthongs (/iaʊ/, /iəʊ/, /uai/ and /uei/), all of them have glides /i/ or /u/ and we set their weights at 1/5 and weights of each of the two remaining component vowels at 2/5.

For the time-reversed speech, the phonetic and phonemic feature matrices were time reversals of the corresponding forward speech (real-words or pseudo-words).

All features were subsequently downsampled to 64 Hz as for the EEGs. For the spectrogram and spectrogram derivatives, this was conducted using the same decimating approach as for EEGs via a 30th-order Hamming-windowed FIR filter. For the phonetic and phonemic features, because they are discrete features, this was done by simple resampling rather than decimation.

2.5 | Reconstruction of EEGs using stimulus features

2.5.1 | The mTRF model

We employed multivariate temporal response functions (mTRFs; Crosse et al., 2016; Di Liberto et al., 2015) to model EEGs using the stimulus features (i.e., forward encoding) via the following formula:

$$r_i(t) = \sum_j \sum_{\tau=0}^{\tau_{\max}} TRF_{ij}(\tau) s_j(t-\tau) + \epsilon_i(t) \quad (1)$$

$r_i(t)$ is the EEG time series at the ith electrode. $s_j(t)$ is the time series of the jth vector of the stimulus features. $TRF_{ij}(t)$ is the time series of the TRF. $\varepsilon_i(t)$ is the error term. τ is the time lag between the EEG and the stimulus feature series and $\tau_{\max}$ is the maximum lag set at 500 ms in the present study. The TRF was estimated by minimising the mean squares of errors. As such, TRF can be obtained via the following matrix formula:

$$TRF_{i\lambda} = \left(\boldsymbol{S}^T\boldsymbol{S} + \lambda \boldsymbol{I}\right)^{-1}\boldsymbol{S}^T r_i \quad (2)$$

$\boldsymbol{S}$ is a matrix comprised of a lagged time series of the stimulus features. $\boldsymbol{r_i}$ is the vector of EEG series at the ith electrode. λ and $\boldsymbol{I}$ denote the ridge regression parameter and an identity matrix, respectively (see Crosse et al., 2016). The ridge regularisation avoided the ill-posed estimation and overfitting.

We used Matlab 2022a in combination with the mTRF Toolbox (Crosse et al., 2016; https://cnspworkshop.net/resources.html) to fit the models for delta- and theta-band EEGs using the stimulus features as described above (spectrogram, phonetics and phonemes and the combination of these features). To avoid the transient effect at the stimulus onset, we excluded the first 250 ms (roughly corresponding to duration of one syllable), hence 2 s period of interest for each trial. The modelling procedures involved cross-validation to tune the λ values to optimise TRFs followed by testing that estimated EEG reconstruction accuracies using the optimised TRFs (Section 2.5.2). Furthermore, because different features could be highly correlated with each other, a partialling procedure was employed to single out unique contributions of individual features (Section 2.5.3). Reconstruction accuracies were then estimated at individual time lags (over the 500 ms range; Section 2.5.4). Finally, we separately computed the reconstruction for the first and second syllables of the disyllabic words in the forward speech to reveal how semantics (real-words vs. pseudo-words) may affect neural tracking (esp. sharpening) at different syllable positions (Section 2.5.5).

2.5.2 | Cross-validation and testing

Artefact-free trials were first divided into training sets and testing sets. Here, for each stimulus type, we partitioned the trials into

five subsets with each subset having the same or similar number of trials (same number of trials was not always guaranteed because the total number of artefact-free trials may not be divisible by five). A training–testing procedure was then run for five times, each of which used one subset as the testing set with the remaining trials as the training set (hence the ratio of trial numbers of training:testing was 4:1). On average, there were 48.6 (12.15), 47.96 (11.99) and 48.28 (12.07) trials for the training (testing in the brackets) for real-words, pseudo-words and time-reversed speech, respectively. EEG reconstruction accuracies were obtained via such procedure and the final accuracy was taken as the average over the five times. This thus made sure that all trials had the equal opportunity to be both training and testing trials, so that biases between individual trials were minimised. Concerns may also be raised that these trial numbers for training may not result in robust neural tracking. To relieve this concern, we conducted additional analyses using different percentages of the number of training trials (60%, 70%, 80%, 90% and 100%) to see how many trials are needed to obtain stable and consistent tracking. Subsequently, we found that the overall patterns (i.e., how reconstruction accuracies differ between stimulus types) are fairly consistent with the current findings even when the number of training trials is as low as 60% of the total number. Stable statistical patterns can be obtained based on >80% of the training trials (i.e., ∼40 trials) with accuracies comparable to previous reports that used lengthy audiobooks as stimuli (Di Liberto et al., 2015; Di Liberto & Lalor, 2017; see Supplementary Materials S1). This therefore in part validates the result robustness based on the current number of training trials (please see Section 4.4 for further explanations and discussions).

During each training–testing procedure, leave-one-out cross-validation was employed to optimise the ridge parameter λ within the training set (Crosse et al., 2016). First, one trial was selected to be left out as a validator with TRFs averaged across the remaining trials computed using the stimulus features and λ with a range of values (10^{-4}, $10^{-3.5}$, 10^{-3}, …, 10^{3}, $10^{3.5}$ and 10^{4}). The TRFs were then tested on the validator trial to obtain Pearson correlations between the predicted EEG and the actual EEG of the validator trial. Second, a different trial was then selected as the validator in the next round of validation and such procedure was repeated until all trials were assigned as validators. The Pearson correlation values were then averaged across all validators and all 32 EEG electrodes. The optimal λ value was identified as the one which yielded the highest correlation (i.e., best model fit for the training set).

The TRFs were then computed using the corresponding optimal λ (formula (2) in Section 2.5.1) and were averaged across all trials within the training set. The averaged TRF was then used to test on the testing trials to estimate EEG reconstruction accuracies as Pearson correlations (Fisher-transformed) between the predicted EEG and the actual EEG. The correlation values were then averaged across all testing trials and the five times of the training–testing procedures.

Furthermore, to test whether reconstruction accuracies were above chance, we computed the shuffled accuracies by randomly permuting the trial correspondence between the stimulus features and the to-be-predicted EEGs for the testing trials (meanwhile ensuring all trials are unmatched). The permutation was repeated 100 times for each partition of training/testing sets, hence 100 shuffled accuracy values for every testing trial (this therefore gives on average > 1000 shuffled accuracies for each partition). This is important because there would be concerns about whether reliable/robust neural tracking by mTRF models can be obtained based on a short stimulus duration (2 s analyses period, see Section 2.5.1) in each utterance (see discussions in Section 4.4). We confirmed the validity of model fitting to obtain reliable neural tracking by showing significantly higher reconstruction accuracies than shuffled accuracies (see results in Section 3.2.1). Besides confirming such validity, shuffled accuracies were also used to obtain normalised accuracies in addition to the original accuracies. Both original and normalised accuracies were used for subsequent mixed-effect regressions (see Section 2.8.2 for details).

2.5.3 | Singling out contributions of individual features

As different stimulus features (spectrogram, phonetics and phonemes) could be highly correlated with each other (Kriegeskorte & Douglas, 2019), a partialling procedure was applied to single out unique contributions of the respective features to predict EEGs. We used the Matlab function *mTRFpartial.m* (Crosse et al., 2021; provided by Dr Aaron Nidiffer). Specifically, a model was fit using the to-be-partialled features (e.g., phonetics and phonemes) via cross-validation in all trials for a given condition (before partitioning trials into training/testing sets). The TRF computed by this model was used to predict the to-be-partialled EEG by the to-be-partialled features in each trial. This predicted EEG was then subtracted from the actual EEG to obtain the residual EEG for each trial. The residual EEGs were then used for model fitting using the target feature (e.g., spectrogram) following the same training–testing procedures described above (see Section 2.5.2). Such approach has been validated by recent research and shown that it is equivalent to the partial correlations that statically control for contributions of the to-be-partialled features (Teoh et al., 2022).

2.5.4 | Predicting EEGs at individual time lags

To further investigate the temporal dynamics of neural tracking, EEG reconstruction accuracies were calculated at individual time lags (τ) across 0–500 ms. The time step was set at 15.6 ms, that is, one cycle of the downsampled frequency for the model fitting (64 Hz). Specifically, at each time step, we used the corresponding TRF (computed via cross-validation in Sections 2.5.2 and 2.5.3) to predict the EEG over a short period of lags centred at that step (across three sampling points, e.g., accuracy at the lag of 250 ms was obtained by EEG prediction using the TRF over lags of 234.4, 250 and 265.6 ms). This was done for both when combining all stimulus features and when specific effects of individual features were singled out.

2.5.5 | Reconstructing EEG at the first and second syllables within disyllabic words

Results for theta-band tracking did not show significant semantic effects (real-words vs. pseudo-words) following the procedure from Sections 2.5.1 to 2.5.4 (see Sections 3.2.2 and 3.2.4). We suspected that this might be because while there should be priming effects within the disyllabic words, such effects would only happen at particular positions of the disyllabic words for the real-words condition (i.e., the first syllables primed the seconds). As such, it may well be that semantic sharpening only took effects at the second, but not the first, syllables. We thus further computed the EEG reconstruction accuracies separately for the first and second syllables within disyllabic words for real-words and pseudo-words. N.B., while every real-word and pseudo-word utterance had the same number of syllables (i.e., nine syllables), the 'first/second syllables' for pseudo-words specifically refer to pseudo-word syllables at the *corresponding* positions of the first/second syllables within disyllabic words in the real-word utterances. Therefore, while the first syllables within disyllabic words are the 1st, 3rd, 5th, and 8th syllables in each real-word utterance (corresponding to the four disyllabic words in each utterance), the 'first syllables' in *pseudo-words* refer to syllables at the corresponding positions (i.e., the 1st, 3rd, 5th, and 8th syllables in each pseudo-word utterance as in real-words); likewise, the second syllables within disyllabic words refer to the 2nd, 4th, 6th and 9th syllables in each real-word and pseudo-word utterance. Specifically, the predicted EEG in each trial, which was obtained following the previous procedures, was used to correlate with the actual EEG not over the entire target period, but over sub-periods at the first and second syllables, respectively (correlations over the first 100 ms after syllable onset which were shown to have the most robust semantic sharpening effect, see Broderick et al. (2019) and Broderick and Lalor (2020). Note that while both the actual and predicted EEGs had been already mean-centred over the entire target period, mean-centring was not duplicated when conducting correlations over these sub-periods. Similar to Section 2.5.4, the reconstruction accuracies were obtained at individual time lags for individual features (with partialling) as well as when all features combined (without partialling). Also, to avoid onset transient effects, the first two syllables of each utterance (correpsonding to the first disyllabic word for each real-word utterance) were not used for such analyses.

2.6 | Relation between alpha-band power and EEG reconstruction accuracies

An alternative explanation for any effect proposed by sharpening and predictive coding is that attention was paid to a specific stimulus type which as a result enhanced neural tracking of that stimulus type. We thus further looked into whether changes in alpha power, which index attentional control during speech perception (O'Sullivan et al., 2019; Wöstmann, Lim, & Obleser, 2017), are related to EEG reconstruction accuracies. We used sliding windows (500 ms long and steps of 100 ms) covering pre-stimulus (a 1-s period before stimulus onset) and the target period of interest (250–2250 ms after stimulus onset). The alpha power in each window was taken as the average log-power of the Fourier spectrum across 8–12 Hz across the parieto-occipital electrodes (Pz, P3, P4, P7, P8, PO3, PO4, Oz, O1, O2, see Figure 2c; referring to Wöstmann, Lim, & Obleser, 2017; O'Sullivan et al., 2019). We calculated the changes in alpha power over the target period relative to the pre-stimulus period. We then tested whether they differed between stimulus types and assessed whether greater negative changes in alpha power (indicating greater attention) are correlated with higher EEG reconstruction accuracies.

2.7 | Fitting with the hierarchical predictive coding framework

To further investigate how the results might fit with the hierarchical Predictive Coding framework (Friston, 2005, 2010), we measured the directed information transfers between prediction errors and sharpened signals across linguistic levels. In the framework (see Figure 7a), 'prediction error units' (brown boxes) feedforward prediction errors to the 'prediction units' (green boxes), followed by feedback of sharpened signals from the prediction units to the prediction error units (Friston, 2005, 2010). The feedforward signals flow from a lower to a higher hierarchical linguistic level (i.e., from acoustic to phonetic level and from phonetic to phonemic level, Figure 7a). The feedback signals, on the other hand, follow the opposite directions to the feedforward signals. In this way, the feedforward flow transfers prediction errors from the lower hierarchical level to the prediction unit at the higher hierarchical level to generate an updated prediction with the sharpened signals; the feedback flow then transfers the sharpened signals (which indicate how predicted the stimulus is) back to the lower-level prediction error unit so that prediction errors are also updated. Such recurrent loops of transfers hence help to support speech perception over time (Sohoglu & Davis, 2016).

The feedforward and feedback flows were computed based on the temporal variations of EEG reconstruction accuracies across individual time lags (see Section 2.5.4). Prediction errors and sharpened signals used to fit the framework were represented by the real-words versus pseudo-words differences in the delta- and theta-band reconstruction accuracies, respectively, for multi-level features (spectrogram, phonetics and phonemes). These were based on the entire analysis period in each utterance, independent of whether statistically significant prediction errors or sharpening effects were actually observed in the previous steps. We applied phase transfer entropy (PTE; Hillebrand et al., 2016; Lobier et al., 2014; Wilmer et al., 2012) that quantifies the directed transfers between prediction errors and sharpened signals. We used the open-accessed Matlab function *PhaseTE_MF.m* (Fraschini & Hillebrand, 2017). Importantly, PTE measures the amount of uncertainty reduced in future values (phase values here) of one signal *Y* by knowing the past values of the other signal *X* given the past values of *Y*. The uncertainty of phase values is measured using Shannon entropy:

$$\begin{aligned}\mathrm{PTE}_{X\to Y} &= H(\theta_{Y,t})|\theta_{Y,t-\delta}) - H(\theta_{Y,t}|\theta_{Y,t-\delta},\theta_{X,t-\delta}) \\ &= H(\theta_{Y,t},\theta_{Y,t-\delta}) - H(\theta_{Y,t-\delta}) - \Big(H(\theta_{Y,t},\theta_{Y,t-\delta},\theta_{X,t-\delta}) \\ &\quad - H(\theta_{Y,t-\delta},\theta_{X,t-\delta})\Big)\end{aligned} \tag{3}$$

$$H(\theta_{X,t}) = -\sum_{i=1}^{N} p_i(\theta_{X,t})\log(p_i(\theta_{X,t})) \tag{4}$$

$$N = 1.87(L-1)^{0.4} \tag{5}$$

In formula (3), θ refers to the Hilbert phase in the time series. δ is the delay between X and Y. Therefore, $\theta_{X,t}$ refers to the phase of X at time t while $\theta_{X,t\text{-}\delta}$ refers to the phase of X in the past with the delay of δ. H refer to Shannon entropy. Formula (4) demonstrates the entropy calculation through phase binning (Hillebrand et al., 2016; Lobier et al., 2014; Wilmer et al., 2012), with p_i as the probability of phase occurrence in the ith bin and N as the total number of bins. N was determined according to formula (5) (Otnes & Enochson, 1972; Pereda et al., 2005), where L is the total number of data samples. Here, we set L as the number of samples of individual time lags (i.e., 32 points over 500 ms) multiplied by the number of electrodes of interest (22 centro-frontal electrodes, see Figure 2c and Section 2.8). The resulted N was 26, hence each bin size was $\pi/13$ (i.e., 2π divided by 26). In order to have enough data to compute PTE, p_i was first measured for all electrodes of interest and was then averaged across electrodes before computing the entropy. This had avoided the scenarios where probability for some phase bins was zero if entropy was computed electrode-wise (which would make entropy unmeasurable).

After obtaining the PTE for the feedforward and feedback flows, we further computed the 'directed PTE' (dPTE; Hillebrand et al., 2016) to quantify the 'net' flows of a particular direction:

$$\mathrm{dPTE} = \frac{\mathrm{PTE}_{\mathrm{feedforward}}}{\mathrm{PTE}_{\mathrm{feedforward}} + \mathrm{PTE}_{\mathrm{feedback}}} \tag{6}$$

dPTE > 0.5 indicates greater feedforward than feedback flows (i.e., net feedforward transfers), while dPTE < 0.5 indicates greater feedback than feedforward flows (net feedback transfers). We measured dPTE at different δ ranging from 15.6 to 250 ms (steps at 15.6 ms, i.e., one cycle of 64 Hz) to study the changes in dominance of feedforward and feedback transfers across delays.

2.8 | Statistical analyses

2.8.1 | Model comparisons to determine the best acoustic representation

We focused on 22 centro-frontal electrodes of interest for statistical analyses (Figure 2c). EEG reconstruction accuracies were averaged across these electrodes. Regions covered by these electrodes should best represent the mTRF model fitting during auditory speech processing (Crosse et al., 2016). This is also proved in our topographic results (see Section 3). As mentioned in Section 2.4, we first determined that spectrogram was the best acoustic representation for the model fitting. This was done by comparing reconstruction accuracies predicted by different acoustic features (see Section 2.4). We conducted bootstrapping (via Matlab R2022a) to quantify the statistical significance (Efron & Tibshirani, 1994). Specifically, for any given within-subject comparison, the data (within-subject differences across participants) were resampled with replacement in each repetition (10,000 repetitions) and a bootstrap distribution was obtained. Following this, a null (H_0) distribution was generated by subtracting the bootstrap distribution from its own mean (so that the distribution mean was zero). p-value was then computed by its definition, that is, the probability of the bootstrap results when they are more extreme than the actual test statistic given the null distribution (Wasserstein & Lazar, 2016):

$$p = 2\min\{Pr(T \geq t|H_0), Pr(T \leq t|H_0)\} \tag{7}$$

where Pr refers to probability, T is the bootstrap result (given the null distribution) and t is the actual test statistic (here the mean value of the within-subject difference). p-value was measured as the proportion of $T \geq t$ (if t is positive) or $T \leq t$ (if t is negative) in the null distribution (hence $\min\{\Pr(T \geq t|H_0), \Pr(T \leq t|H_0)\}$) and was multiplied by two for the two-tailed test. p needs to be <.05 in order to reject the null hypothesis.

2.8.2 | Linear mixed-effect regressions

After determining spectrogram as the best acoustic representation, we predicted the EEGs using spectrogram, phonetic and phonemic features and singled out contributions of individual features (see Section 2.5.3). We then conducted linear mixed-effect regressions to quantify whether the models based on these individual features differed significantly between delta- and theta-band EEGs. Specifically, we used EEG reconstruction accuracy (averaged across the centro-frontal electrodes, Figure 2c) as the dependent variable, Frequency (delta vs. theta), Feature (spectrogram, phonetics and phonemes) and Stimulus Type (real-words, pseudo-words and time-reversed speech) as the fixed-effect factors, and Participant as the random-effect factor (a random intercept for each participant). In additional, including random slopes for within-subject factors is shown to lead to better model fitting (Barr et al., 2013). We therefore further fitted a random slope for Frequency and Stimulus Type, respectively, however, excluded fitting a random slope for Feature because this was shown to result in model singularity. The regression was performed using RStudio (Build 554; RStudio BPC) using the function *lmer* (based on packages *lme4* and *lmerTest*; Bates et al., 2015; Kuznetsova et al., 2017). The following formula was used:

$$ReconstrAcc \sim Freq * Fea * StimType + (Freq + StimType|Participant) \tag{8}$$

ReconstrAcc, *Freq*, *Fea* and *StimType* are abbreviations of Reconstruction Accuracy, Frequency, Feature and Stimulus Type,

respectively. The term *Freq*Fea*StimType* in RStudio includes all factorial terms (main effects and interactions). The fixed-effect factors were all sum-coded before the formula was applied. The model was fitted by restricted maximum likelihood (REML). The degrees of freedom and *p* values for fixed-effect factors were estimated via Satterthwaite approximation. Furthermore, as we were also specifically interested in whether semantic contents (i.e., real-words vs. pseudo-words) would drive the differences in neural tracking at different frequency bands, we conducted additional analyses with Stimulus Type that only included real-words and pseudo-words.

Moreover, to confirm the outcome validity, the regressions were also conducted for reconstruction accuracies normalised based on the chance level (i.e., the null/shuffled accuracies, see Section 2.5.2). The normalised accuracies are computed as (i) differences between the original and shuffled accuracies; and (ii) *z*-scores (electrode-wise and for every testing trial) obtained by subtracting the shuffled mean and then divided by the standard deviation of the shuffled accuracies across testing trials within every given testing/training partition; see Section 2.5.2). We subsequently showed that using original and normalised accuracies resulted in similar statistical outcomes that led to the same conclusions (see Section 3.2.2 and Supplementary Materials S1).

2.8.3 | Subsequent analyses using bootstrapping

Following the linear mixed-effect regressions, we followed the same bootstrapping procedure to conduct statistical comparisons for the planned analyses: accuracies between stimulus types for each partialled speech feature (Section 2.5.3) and across individual time lags (Sections 2.5.4 and 2.5.5), changes in alpha power (Section 2.6) and directed PTE across time delays (Section 2.7). False discovery rate (FDR) corrections on *p* values were further applied according to multiple comparisons (see Section 3 for more details when reporting the statistical findings).

3 | RESULTS

3.1 | Behavioural results

The behavioural task in each trial was to judge whether the cue sound was present in the target utterance in order keep participants' attention throughout the target period. Accuracies and reaction times (interval between the appearance of the question mark and the button-press in each trial) were recorded. Accuracies for the real-words, pseudo-words and time-reversed speech were 98.1 ± 0.4%, 97.6 ± 0.5% and 73.1 ± 2.7%, respectively (mean ± standard error across participants). Reaction times for real-words, pseudo-words and time-reversed utterances were 502.3 ± 47.1, 484.4 ± 40.6 and 691.5 ± 67.8 ms, respectively. We applied the bootstrapping as proposed for the neural analyses (see Section 2.8) and found that all three accuracies were significantly greater than chance (50%; $p < .0001$).

TABLE 1 Behavioural results showing the accuracies and reaction times of the three stimulus types (mean ± standard error across participants).

Stimulus type/comparisons	Accuracy	Reaction time
Real-words	98.1 ± 0.4%	502.3 ± 47.1 ms
Pseudo-words	97.6 ± 0.5%	484.4 ± 40.6 ms
Time-reversed	73.1 ± 2.7%	691.5 ± 67.8 ms
Real-words versus pseudo-words	$p = .2196$	$p = .0952$
Real-words versus time-reversed	***$p < .0001$	***$p < .0001$
Pseudo-words versus time-reversed	***$p < .0001$	***$p < .0001$

Note: *p* values (FDR corrected across stimulus types) indicate whether accuracy and reaction time differed significantly between stimulus types. Significant *p* values are indicated in bold.
***$p < .0001$.

Accuracy and reaction time were compared with *p* values FDR corrected according to the number of stimulus types (i.e., three). The accuracy was significantly greater for real-words and pseudo-words than time-reversed speech (both $p < .0001$) but did not differ between real-words and pseudo-words ($p = .2196$). Likewise, the reaction time was significantly shorter for real-words and pseudo-words compared to time-reversed speech (both $p < .0001$) but did not differ between real-words and pseudo-words ($p = .0952$). Table 1 summarises the accuracies and reaction times across the stimulus types as well as the *p* statistics.

3.2 | Model fitting using mTRF

3.2.1 | Determining the best acoustic representation for model fitting

Delta- and theta-band EEGs were first modelled using acoustic features of delta/theta envelopes (fluctuated at the corresponding EEG frequencies), spectrogram and spectrogram derivatives. As described in Section 2.4.1, we also looked into the model fitting using spectrotemporal modulations. This is because previous neurophysiological and fMRI studies have shown that cortical responses can be well-modelled using spectrotemporal modulations (Daube et al., 2019; Pasley et al., 2012; Santoro et al., 2014; Sohoglu & Davis, 2020). However, using spectrotemporal modulations failed to show significantly greater EEG reconstruction accuracy than shuffled accuracy for either delta- or theta-band tracking for any stimulus type. This may be because spectrotemporal representations could be generic and different stimuli may share common/homogeneous spectrotemporal profiles, especially, when all stimuli were spoken by the same individual [which is the case in the present study and all utterances had the almost the same syllable rate that may further strengthen the similarity of spectrotemporal modulation properties; also see relevant

discussions by Sohoglu and Davis (2020)]. Therefore, we did not consider such feature for further analyses.

EEG reconstruction accuracies and the corresponding shuffled/null accuracies (averaged across centro-frontal electrodes of interest, see Figure 2c) were shown as Figure 3a. To determine the best acoustic representation, we averaged accuracies across frequency bands (delta and theta) and stimulus types (Figure 3b) and assessed which feature resulted in the highest accuracy. The results showed that accuracies were all significantly higher than shuffled accuracies (for delta/theta envelopes: $p = .0106$; for spectrogram derivatives and spectrogram: $p < .001$), confirming the reliability of neural tracking using mTRF models based on the short stimulus analyses duration (2 s; see Section 2.5.1). The accuracy reconstructed using spectrogram was the highest and significantly greater than using the spectrogram derivatives ($p = .0090$) and delta/theta envelopes ($p < .0001$). Combination of spectrogram and derivatives improved the accuracy compared to using derivatives alone but did not yield numerically higher accuracy than spectrogram alone (not shown in the figure). Therefore, we concluded that spectrogram was the best acoustic representation and thus used it as the acoustic feature in subsequent analyses. Furthermore, as expected, we showed that adding phonetic and phonemic features based on spectrogram significantly improved reconstruction accuracy ($p = .0326$; Figure 3b), confirming the superiority of including higher-level features. Also, accuracies for encoding spectrogram and combined spectrogram, phonetic and phonemic features are significantly greater than the corresponding shuffled accuracies ($p < .0001$) as expected. The p values were all computed using bootstrapping (see Section 2.8.1) and were not corrected.

3.2.2 | Neural tracking across frequency bands and stimulus types

We then followed the partialling approach (see Section 2.5.2) to single out the unique contributions of individual features. After that, linear mixed-effect regressions were conducted using reconstruction accuracy (averaged across the centro-frontal electrodes) as the dependant variable, Frequency, Feature and Stimulus Type as the fixed-effect factors, Participant (random intercept), Frequency and Stimulus Type (random slopes) as the random-effect factors (see Section 2.8.2). These were conducted when all stimulus types were included and when only real-words and pseudo-words were included. The statistical results are shown in Table 2. When all stimulus types were included, there was a significant [Frequency × Stimulus Type] interaction ($F(1, 294.998) = 4.748$, $p = .0301$) and main effects of Frequency

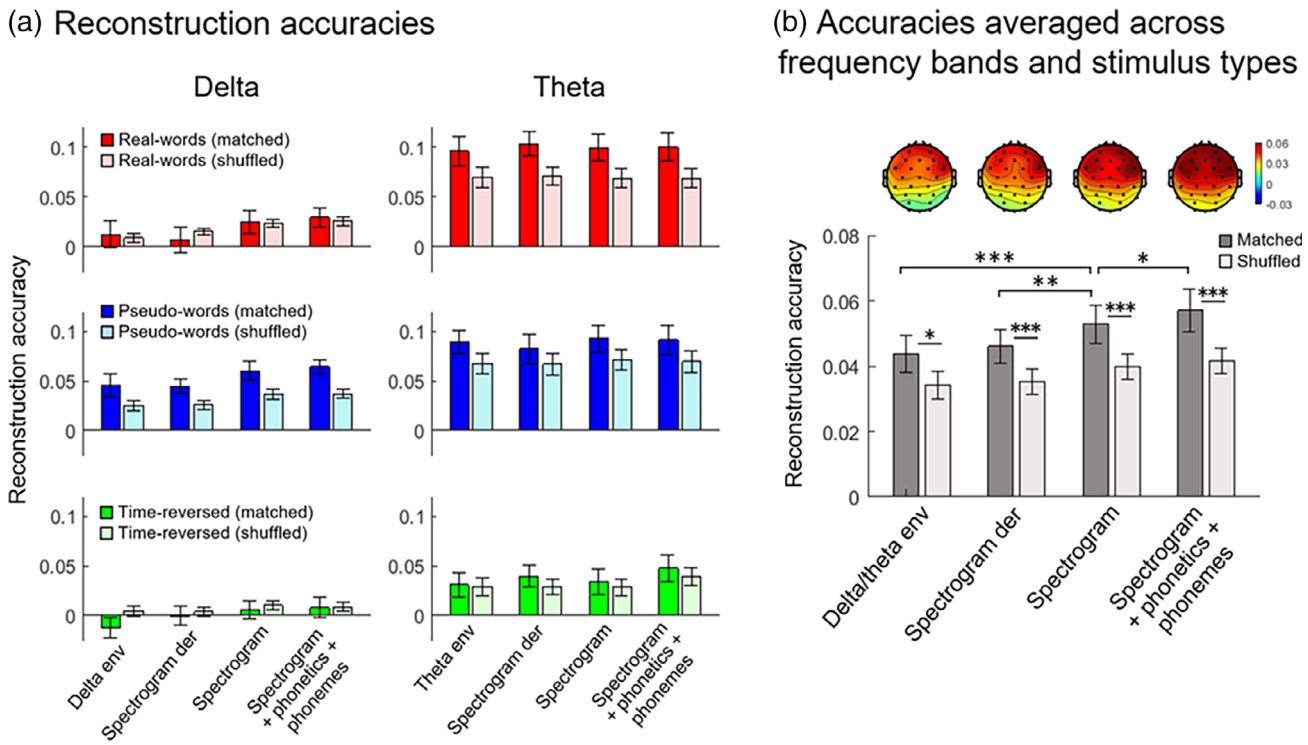

FIGURE 3 Reconstruction accuracies (averaged across centro-frontal electrodes shown in Figure 2c) for features of delta/theta (same frequencies as for EEG) envelopes, spectrogram derivatives, spectrogram and spectrogram plus higher-level (phonetic and phonemic) features. (a) Accuracies across features, frequency bands (delta and theta) and stimulus types (real-words, pseudo-words and time-reversed speech). The original (with matched correspondence of trials) and null (with shuffled and unmatched correspondence of trials) accuracies are indicated by darker and lighter colours, respectively. (b) Accuracies averaged across frequency bands and stimulus types (corresponding topoplots on the top). Spectrogram yielded the highest reconstruction accuracy amongst the acoustic features. As expected, adding phonetic and phonemic features resulted in significant improvement in accuracy. *$p < .05$; **$p < .01$; ***$p < .001$, all p values are *uncorrected*.

TABLE 2 Statistical results for the linear mixed-effect regressions.

DV	Fixed-effect factors/interactions	df1	df2	*F*	*p*
Reconstruction accuracy (Stimulus Type includes real-words, pseudo-words and time-reversed speech)	Frequency × Feature × Stimulus Type	1	294.998	<0.001	.9935
	Frequency × Feature	1	294.998	0.118	.7313
	Frequency × Stimulus Type	1	294.998	4.748	**.0301***
	Feature × Stimulus Type	1	294.998	3.808	.0520
	Frequency	1	18.998	20.590	**.0002*****
	Feature	1	294.998	2.593	.1084
	Stimulus Type	1	19.003	14.024	**.0014****
Reconstruction accuracy (Stimulus Type includes real-words and pseudo-words)	Frequency × Feature × Stimulus Type	1	175.001	0.140	.7089
	Frequency × Feature	1	175.001	0.266	.6064
	Frequency × Stimulus Type	1	175.001	13.513	**.0003*****
	Feature × Stimulus Type	1	175.001	0.247	.6201
	Frequency	1	19.000	15.416	**.0009*****
	Feature	1	175.001	9.533	**.0023****
	Stimulus Type	1	18.999	2.333	.1432

Note: These were conducted when including all stimulus types (first part) and when only including real-words and pseudo-words (second part). DV, *df*, *F* and *p* refer to dependent variable, degree of freedom, *F* and *p* values, respectively. Degrees of freedom and *p* values for fixed-effect factors are estimated via Satterthwaite approximation. Significant *p* values (<.05) are indicated in bold.
*$p < .05$; **$p < .01$; ***$p < .001$.

($F(1, 18.998) = 20.590$, $p = .0002$) and Stimulus Type ($F(1, 19.003) = 14.024$, $p = .0014$). When only real-words and pseudo-words were included, there was a significant [Frequency × Stimulus Type] interaction ($F(1, 175.001) = 13.513$, $p = .0003$) and main effects of Frequency ($F(1, 19) = 15.416$, $p = .0009$) and Feature ($F(1, 175.001) = 9.533$, $p = .0023$). Here, the significant [Frequency × Stimulus Type] interactions are of particular interest, because they indicate that neural tracking of different stimulus types differed between delta and theta bands. This is particularly the case when only real-words and pseudo-words were included (with a much lower *p*-value).

Following the significant [Frequency × Stimulus Type] interactions, we analysed how reconstruction accuracies (averaged across centro-frontal electrodes) differed between stimulus types for delta- and theta-band accuracies separately. Bootstrapping was conducted to compute *p* values for evaluations of significance (FDR corrected according to the three stimulus types for each frequency band and feature, Figure 4a). Statistical results are as follows: (1) When combining all features (spectrogram + phonetics + phonemes) in the mTRF models, pseudo-words had significantly greater delta-band accuracy than real-words ($p = .0027$) and time-reversed speech ($p = .0003$). No significant difference in delta-band accuracy was found between real-words and pseudo-words ($p = .1292$). Both real-words and pseudo-words had significantly greater theta-band accuracies than time-reversed speech ($p = .0030$ and .0378, respectively). No significant difference in theta-band accuracy was found between real-words and pseudo-words ($p = .3304$). (2) During spectrogram encoding (when contributions of phonetics and phonemes were partialled out), real-words had significant greater delta- and theta-band accuracies than time-reversed speech (delta: $p = .0003$; theta: $p < .0001$). Pseudo-words also had significant greater delta- and theta-band accuracies than time-reversed speech (delta: $p < .0001$; theta: $p < .0001$). No significant differences in either delta- or theta-band accuracies between real-words and pseudo-words (delta: $p = .0890$; theta: $p = .9096$). (3) During phonetic encoding (when contributions of spectrogram and phonemes were partialled out), pseudo-words had significantly greater delta-band accuracy than real-words ($p < .0001$) and time-reversed speech ($p < .0001$). No significant difference in delta-band accuracy was found between real-words and time-reversed speech ($p = .3488$). No significant differences in theta-band accuracy were found between stimulus types (all $p > .1$). (4) During phonemic encoding (when contributions of spectrogram and phonetics were partialled out), pseudo-words had significantly greater delta-band accuracy than time-reversed speech ($p = .0024$). No significant differences in delta-band accuracies were found between real-words and pseudo-words ($p = .0639$) or between real-words and time-reversed speech ($p = .3840$). Real-words had significantly greater theta-band accuracy than time-reversed speech ($p = .0114$). No significant differences in theta-band accuracy were found between real-words and pseudo-words ($p = .4118$) or between pseudo-words and time-reversed speech ($p = .2439$). Table 3 summarises the reconstruction accuracies across the stimulus types and features and the *p* statistics.

In addition, electrode-wise comparisons were conducted between stimulus types for delta- and theta-band accuracies. *p* values were computed using bootstrapping for all electrodes and were FDR-corrected according to the number stimulus types (i.e., three) *and* electrodes (i.e., 32) for each frequency band and feature (Figure 4b). The effects are highly consistent with the results shown in Figure 4a. When combination of all features were encoded (spectrogram + phonetics + phonemes), delta-band accuracy was significantly greater for pseudo-words than real-words and time-reversed speech over

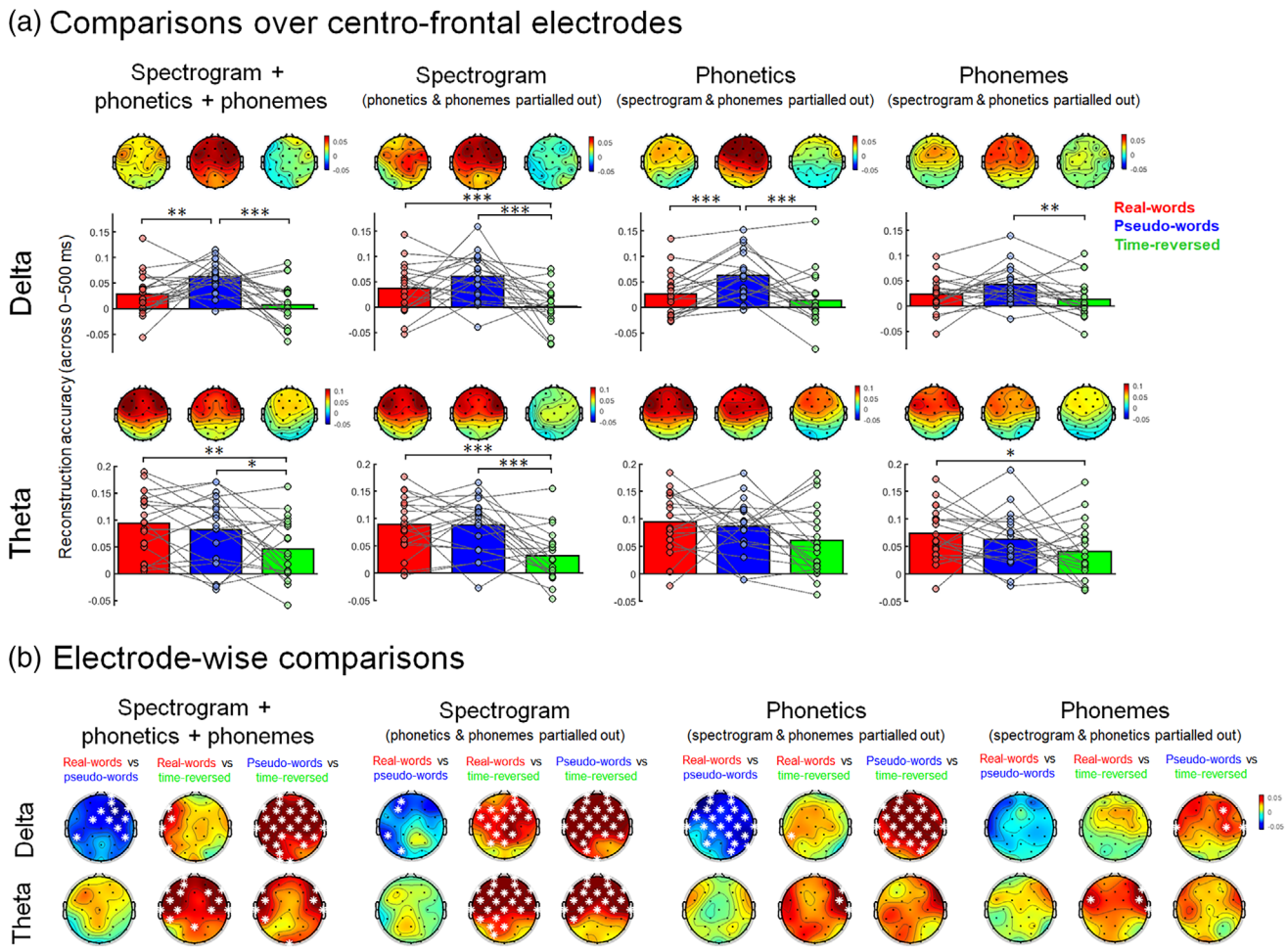

FIGURE 4 Statistical comparisons of reconstruction accuracies between stimulus types for all three features combined (spectrogram + phonetics + phonemes) and for all individual features (after partialling out contributions of any other two features). (a) Delta- (upper panels) and theta-band (lower panels) accuracies showing individual participant values (scattered dots) averaged across centro-frontal electrodes of interest (see Figure 2c). Lines connecting between two given dots indicate these dots came from the same participant according to the within-subject design. Bar magnitudes indicate the mean values across participants. Corresponding topoplots are shown right above the bar graphs. *p* values are FDR corrected according to the three stimulus types for each frequency band and feature. **p* < .05; ***p* < .01; ****p* < .001. (b) Electrode-wise comparisons between stimulus types (delta: upper panels; theta: lower panels). The white asterisks over corresponding electrodes indicate significance differences (*p* < .05, FDR corrected according to the three stimulus types and all 32 electrodes for each frequency band and feature).

multiple centro-frontal and parietal electrodes but only a few significant electrodes on the left hemisphere for real-words versus time-reversed speech (the first upper panel); theta-band accuracy was greater for forward (real-words and pseudo-words) than time-reversed speech over multiple temporo-frontal electrodes but did not differ between real-words and pseudo-words (the first lower panel). For individual features, greater delta-band accuracies for pseudo-words than real-words are most evident during phonetic encoding (the third upper panel) and greater theta-band accuracies for forward than time-reversed speech are most evident during spectrogram encoding (the second lower panel).

Similar statistical results were also obtained based on normalised reconstruction accuracies (see Section 2.8.2) that lead to the same conclusion as the non-normalised original accuracies (see Supplementary Materials S1). We here focus on the non-normalised accuracies. This is because all trials shared common rhythmic speech properties (all utterances were produced at the same syllable rate at ~4 Hz with all syllables having a similar duration; see Section 2.2). These shared properties across trials might contribute to similar patterns for shuffled accuracies as for the non-normalised accuracies as a result of sharpening/predictive coding [esp. during encoding spectrogram which is shaped by rhythmicity; similar concerns as in Sohoglu and Davis (2020)]. In this case, the normalisation may, to a certain extent, smear the effects of sharpening and predictive coding, so it may be more appropriate to base our result interpretations on the non-normalised accuracies. Nonetheless, we also emphasise that it is reassuring that the normalised and non-normalised accuracies showed similar statistical outcomes.

TABLE 3 EEG reconstruction accuracies (mean ± standard error) for all three features combined (spectrogram + phonetics + phonemes) and individual features (after partialling out contributions of any other two features) across frequency ranges (delta and theta) and the three stimulus types.

Frequency	Stimulus type/comparisons	All features	Spectrogram	Phonetics	Phonemes
Delta	Real-words	0.028 ± 0.009	0.037 ± 0.011	0.026 ± 0.010	0.023 ± 0.008
	Pseudo-words	0.063 ± 0.007	0.061 ± 0.010	0.063 ± 0.010	0.043 ± 0.008
	Time-reversed	0.008 ± 0.010	0.002 ± 0.009	0.014 ± 0.011	0.014 ± 0.009
	Real-words versus pseudo-words	**$p = .0027$	$p = .0890$	***$p < .0001$	$p = .0639$
	Real-words versus time-reversed	$p = .1292$	***$p = .0003$	$p = .3488$	$p = .3840$
	Pseudo-words versus time-reversed	***$p = .0003$	***$p < .0001$	***$p < .0001$	**$p = .0024$
Theta	Real-words	0.094 ± 0.013	0.090 ± 0.011	0.095 ± 0.012	0.075 ± 0.011
	Pseudo-words	0.083 ± 0.014	0.089 ± 0.011	0.087 ± 0.011	0.063 ± 0.011
	Time-reversed	0.046 ± 0.013	0.032 ± 0.011	0.061 ± 0.014	0.041 ± 0.012
	Real-words versus pseudo-words	$p = .3304$	$p = .9096$	$p = .6146$	$p = .4118$
	Real-words versus time-reversed	**$p = .0030$	***$p < .0001$	$p = .1311$	*$p = .0114$
	Pseudo-words versus time-reversed	***$p = .0378$	***$p < .0001$	$p = .1311$	$p = .2439$

Note: p values (FDR corrected across stimulus types) indicate whether reconstruction accuracies differed significantly between stimulus types. Significant p values (<.05) are in bold.
*$p < .05$; **$p < .01$; ***$p < .0001$.

3.2.3 | Relation between alpha power and neural tracking

We examine whether attention/listening effort may modulate the current findings of sharpening and predictive coding by measuring the changes in alpha power relative to the pre-stimulus baseline (averaged over parieto-occipital electrodes, see Section 2.6). Alpha power indices attentional control during speech perception (greater negative change for greater attention hence listening effort; O'Sullivan et al., 2019; Wöstmann, Lim, & Obleser, 2017). If neural sharpening was driven by attention/listening effort, we should anticipate greater negative change for real-words than pseudo-words and for forward than time-reversed speech. If predictive coding was driven by attention/listening effort, we should anticipate greater negative change for pseudo-words than real-words and time-reversed speech.

Statistical significances were assessed using bootstrapping (see Section 2.8) with FDR correction according to the number of stimulus types (i.e., three). We found significantly greater negative changes in alpha power for time-reversed speech than for real-words ($p = .0087$) and pseudo-words ($p = .0087$; Figure 5a). This indicates that participants paid greatest attention to time-reversed speech, which is plausible because of the greatest task difficulty (also reflected by the behavioural results, see Section 3.1). No significant difference was found between real-words and pseudo-words ($p = .4282$). These results are *not* consistent with any anticipation had sharpening/predictive coding has been driven by attention/listening effort. As there is a trend that pseudo-words had greater negative change than real-words (despite no significant difference), it may still be possible that greater delta-band tracking for pseudo-words is related to such change. We thus conducted an additional correlation between the difference in change in alpha power (pseudo-words vs. real-words) and the difference in delta-band reconstruction accuracy (also pseudo-words vs. real-words; combining all features, i.e., spectrogram + phonetics + phonemes). No significant correlation was found ($r = .3312$, $p = .1537$, Figure 5b; N.B., if greater tracking in pseudo-words was related to greater listening effort, a significant negative correlation should be found). Therefore, we found no evidence that the current findings of sharpening or predictive coding can be plausibly explained by attention/listening effort.

3.2.4 | Neural tracking at individual time lags

EEG reconstruction accuracies were further compared between stimulus types across individual time lags between EEG and speech features to see *when* sharpening/prediction errors were generated. Statistical significances were determined by bootstrapping and were FDR corrected according to the number of stimulus types (i.e., three) and the length of time lags (32 points from 15.6 to 500 ms; see Section 2.8). The results are shown in Figure 6a. Statistical significance is indicated by dark and light brown lines.

For the delta-band (upper panels of Figure 6a), we consistently found significantly higher accuracies for pseudo-words than real-words and time-reversed speech. However, these effects occurred at different time lags during encoding of different speech features. Specifically, when all features were combined, delta-band accuracy was significantly higher for pseudo-words than real-words at both early (16–78 ms; numbers after the decimal point are rounded up/down hereafter) and late time stages (434–500 ms). Such effect occurred briefly at mid stages of 218–250 ms during spectrogram encoding (phonetics and phonemes partialled out); across the early (16–94 ms), mid (218–265 ms) and late stages (406–452 ms) during phonetic

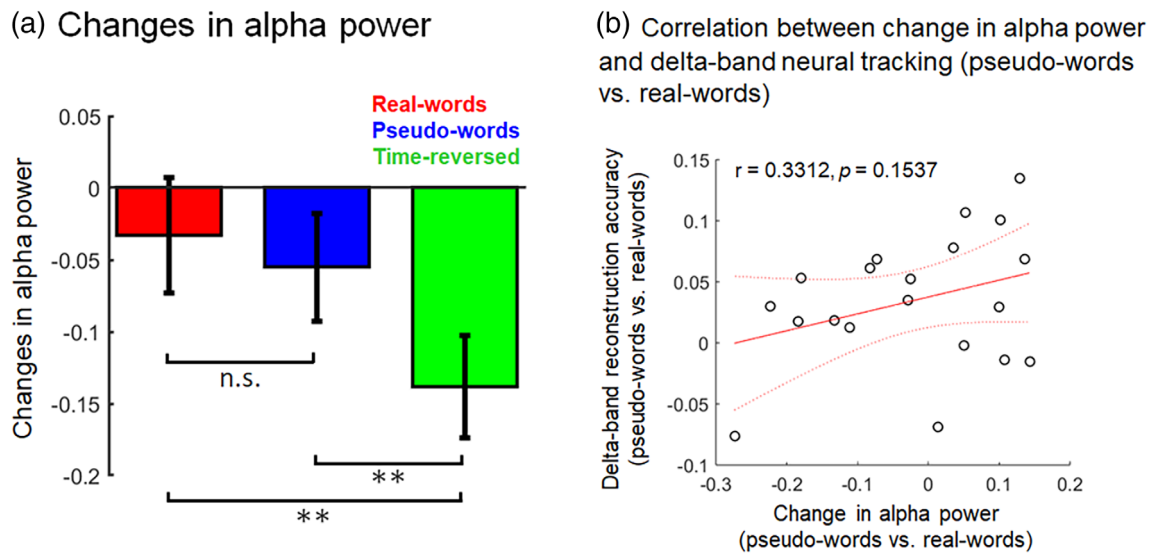

FIGURE 5 Changes in alpha power averaged over parieto-occipital electrodes. (a) Results for changes in alpha power across stimulus types. Significantly greater negative change was found for time-reversed speech than real-words and pseudo-words. n.s., non-significant; $^{**}p < .01$; p values are FDR corrected. Error bars indicate standard errors. (b) Correlation between the change in alpha power and delta-band reconstruction accuracy (pseudo-words vs. real-words). No significant correlation was found.

encoding (spectrogram and phonemes partialled out); and at late stages only (406–468 ms) during phonemic encoding (spectrogram and phonetics partialled out). Significantly higher delta-band accuracies for pseudo-words than time-reversed speech generally occurred across the early, mid and later stages for all features. Consistent with the results in Section 3.2.2, we also found significantly higher delta-band accuracy for real-words than time-reversed speech during spectrogram encoding (78–172 ms). For the theta-band (lower panels of Figure 6a), we consistently found significantly higher accuracies for real-words and pseudo-words than time-reversed speech (spanning over the 500 ms lags except for relatively early stages <250 ms during phonemic encoding), but no significant differences between real-words and pseudo-words.

Following the [Frequency × Stimulus Type] interactions shown in Section 3.2.2, we further assessed such interactions at individual time lags (for Stimulus Type that only includes real-words and pseudo-words that we are particularly interested in). We followed the same bootstrapping approach to compute p values (FDR-corrected according to the total length of time lags) for comparing the real-words versus pseudo-words difference between delta and theta bands. Significances are indicated by purple lines in Figure 6a. We found significant interactions when encoding combination of all features at early stages (16–94 ms) and during phonetic and phonemic encoding at both early (phonetics: 47–94 ms; phonemes: at 62 ms) and late stages (phonetics: 468–500 ms; phonemes: 406–452 ms). These results further show that neural tracking of speech differed at delta- and theta-bands and these differences started at early encoding stages within 100 ms after feature onsets. Furthermore, this occurred mainly during encoding higher-level phonetic and phonemic rather than lower-level acoustic (spectrogram) features.

These results did not find the anticipated semantic sharpening effect (greater theta-band tracking for real-words than pseudo-words) and we suspected this might be because sharpening only occurred at the second syllables of the disyllabic words. Therefore, we further computed theta-band accuracies over the periods of the first and second syllables separately (see Section 2.5.5). Statistical significances were computed by comparing real-words with pseudo-words across time lags using bootstrapping. p values were FDR corrected according to the two types of comparisons (real-words vs. pseudo-words for the first and the second syllables) and the total lengths of time lags. We found significantly higher theta-band accuracies for real-words than pseudo-words for the second syllables when encoding all features combined (Figure 6b; patterns for encoding of individual features are not shown here due to the lack of significant effects). This occurred within 200 ms after feature onsets (16–31 and 125–141 ms; the left panel of Figure 6b). Electrode-wise comparisons showed that this effect occurred over multiple frontal electrodes (mostly in the left hemisphere; p values were FDR corrected according to all 32 electrodes; the right panel of Figure 6b). On the other hand, we did not find any significant real-words versus pseudo-words difference in theta-band accuracy for the first syllables. The results thus indicate that semantic sharpening did exist for theta-band tracking. An additional observation is that theta-band tracking was seemingly greater for the first compared to the second syllables regardless whether they belong to real-words or pseudo-words. This may be because neural tracking was greater when the tracked speech were in temporally earlier positions (first vs. second syllables).

3.3 | Feedforward and feedback transfers in the hierarchical predictive coding framework

Net information flows were computed using dPTE to indicate whether and when feedforward and feedback information flows

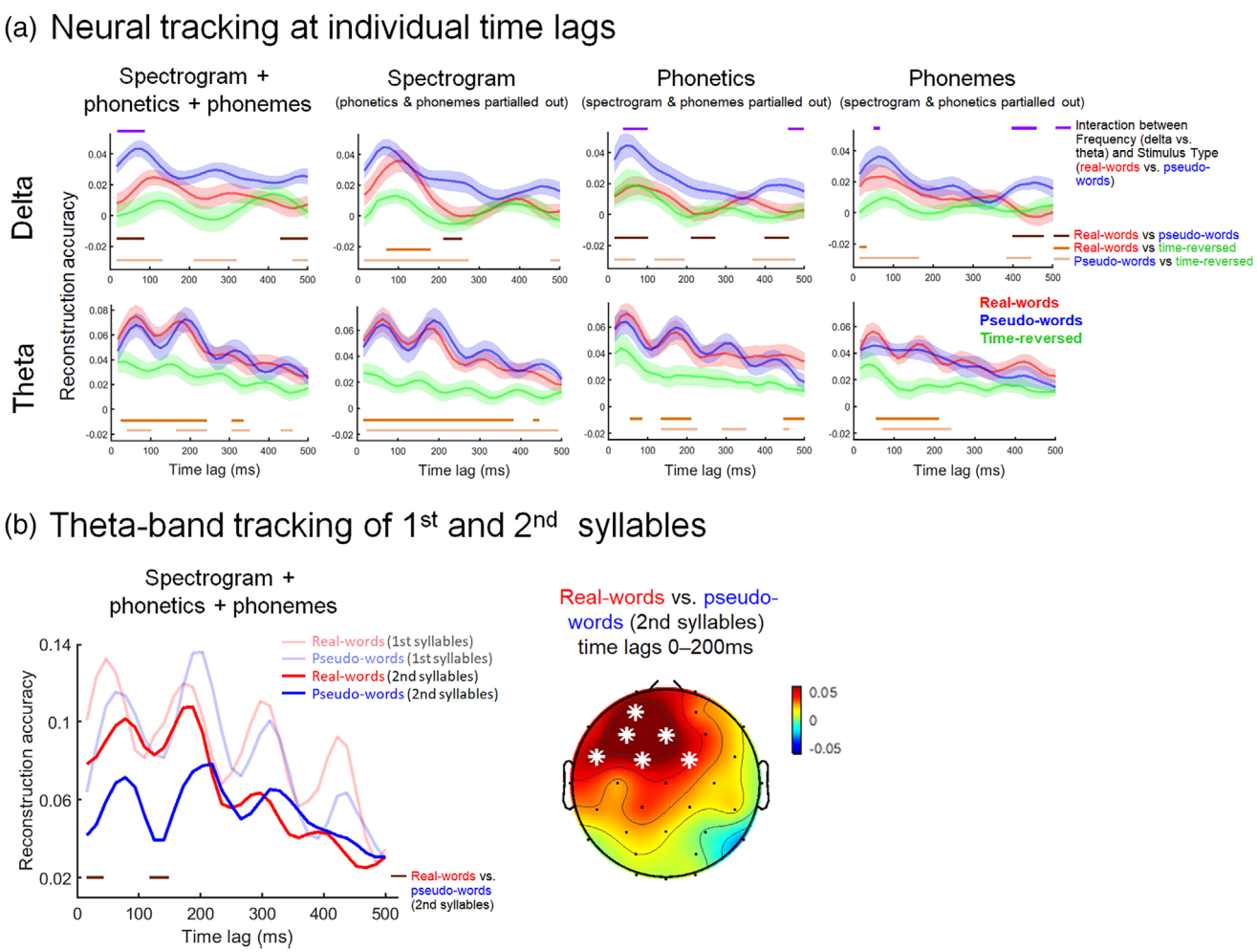

FIGURE 6 Statistical comparisons of reconstruction accuracies across stimulus types at individual time lags (16–500 ms). (a) Delta- (upper panels) and theta-band (lower panels) accuracies across individual time lags (averaged across the centro-frontal electrodes, see Section 2.8.1 and Figure 2c). Shaded areas indicate the ranges of standard errors from the means. Horizontal lines at the bottom of each graph indicate the significant differences between stimulus types (from darker to lighter browns: real-words vs. pseudo-words, real-words vs. time-reversed speech, and pseudo-words vs. time-reversed speech, respectively; $p < .05$, FDR corrected according to the three stimulus types and total length of time lags for each frequency band and feature). Horizontal lines in purple at the top indicate significant interactions between Frequency (delta vs. theta) and Stimulus Type (real-words vs. pseudo-words; $p < .05$, FDR corrected according to the total lengths of lags). (b) Theta-band accuracies (real-words vs. pseudo-words) for the first and second syllables of disyllabic words. The '1st/2nd syllables' for pseudo-words are specifically referred to pseudo-word syllables at the *corresponding* positions of the first/second syllables within disyllabic words in the real-word utterances (see Section 2.5.5). *Left panel*: accuracies over individual time lags for the first and second syllables (averaged across the centro-frontal electrodes in Figure 2c). The horizontal lines at the bottom indicate significant real-words versus pseudo-words differences for the second syllables ($p < .05$, FDR corrected according to the two comparisons and the total lengths of time lags). *Right panel*: topoplot showing the real-words versus pseudo-words differences at the second syllables (averaged over the first 200 ms time lags after feature onsets). The white asterisks at corresponding electrodes indicate significant differences ($p < .05$, FDR corrected according to all 32 electrodes).

dominated the transfers between prediction errors and sharpened signals (see Section 2.7) in the Predictive Coding framework (Friston, 2005, 2010; Figure 7a). The framework proposes that prediction errors (real-words vs. pseudo-words difference in delta-band accuracies) are feedforwarded by the 'prediction error units' to the 'prediction units', followed by feedback of sharpened signals (real-words vs. pseudo-words difference in theta-band accuracies) from the prediction units to the prediction error units. We focused on transfers across linguistic levels (i.e., information feedforwarded from acoustic to phonetic levels followed by feedback from phonetic to acoustic levels, and information feedforwarded from phonetic to phonemic levels followed by feedback from phonemic to phonetic levels). Statistical significances of dPTE were detected via bootstrapping (see Section 2.8) with FDR correction according to the number time delays (15.6–250 ms). The results show significant net feedforward transfers from prediction errors to the sharpened signals (dPTE > 0.5) that occurred mostly within 100 ms delays (indicated by horizontal lines in brown, Figure 7b). This was followed by significant net feedback transfers from sharpened signals to prediction errors (dPTE < 0.5) that occurred at ~150–200 ms (from

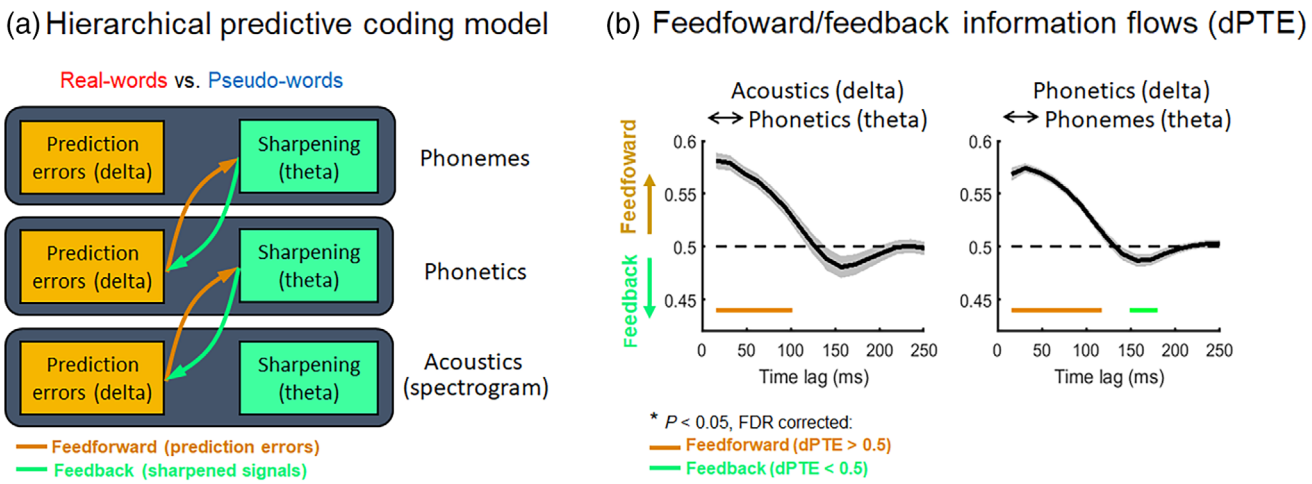

FIGURE 7 Fitting the current findings with hierarchical Predictive Coding framework (Friston, 2005, 2010). (a) Schematic illustration of the framework. Feedforward information for prediction errors are transferred from the 'prediction error units' at lower linguistic levels (brown boxes) to the 'prediction units' at higher linguistic levels (green boxes), followed by feedback of sharpened signals from the prediction units at higher linguistic levels to the prediction error units at lower linguistic levels. (b) Net feedforward (dPTE > 0.5) and feedback (dPTE < 0.5) transfers across time delays (*left*: transfers between acoustic and phonetic levels; *right*: transfers between phonetic and phonemic levels). Net feedforward transfers occurred followed by net feedback transfers (significances indicated by horizontal lines at the bottom in brown and green, respectively). *p* values are FDR corrected according to the lengths of time delays (15.6–250 ms).

phonemic to phonetic levels, indicated by horizontal line in green, Figure 7b).

4 | DISCUSSION

The current study hypothesised that delta- and theta-band neural tracking of multi-level speech features play the roles of predictive coding and neural sharpening, respectively. We examined neural tracking of acoustic (spectrogram), phonetic and phonemic features following partialling procedures so that the tracking is unique for individual features. Our results are consistent with this hypothesis. To the best of our knowledge, this study is the first to illustrate the distinctive roles of delta and theta bands for neural sharpening and predictive coding during spoken language processing. We will discuss our results for multi-level speech encoding (Section 4.1) and interpret the findings of the distinctive neural tracking at delta and theta bands for sharpening and predictive coding (Section 4.2). We will then discuss the early occurrence of sharpening and prediction errors and how our results fit with the hierarchical predictive coding framework (Section 4.3). We will finally discuss possible concerns and how future work may consolidate our current findings (Section 4.4).

4.1 | Neural tracking of multi-level speech features beyond envelopes at the corresponding frequencies of EEG

The current study investigated delta- and theta-band tracking of multi-level speech features from acoustic to phonetic and phonemic features. Traditionally, research has focused on tracking of slowly varying envelopes at the corresponding frequencies of the neural signals (e.g., Ahissar et al., 2001; Etard & Reichenbach, 2019; Mai et al., 2016; Peelle et al., 2013). Interestingly, our results showed that the best acoustic representation for tracking was the spectrogram in which envelopes were low-passed at 30 Hz rather than envelopes bandpass filtered at the corresponding delta/theta frequencies (Figure 3b). This indicates that low-frequency neural signals encode not only speech properties fluctuating at the same frequencies, but also components at other frequencies, which could also be important. This is compatible with findings showing that envelopes fluctuating at either delta or theta band alone are inadequate for speech intelligibility (Arai et al., 1999; Mai, 2014) and components with frequencies higher than delta-theta can also make significant contributions (Shannon et al., 1995; Xu & Pfingst, 2008). Furthermore, we confirmed that, besides acoustic (spectrogram) features, adding higher-level phonetic and phonemic features into the models significantly improved EEG reconstruction accuracies (Figure 3b), consistent with previous studies (Di Liberto et al., 2015; Di Liberto & Lalor, 2017). The current results thus stress that the neural sharpening and predictive coding occur not only for the processing of envelope cues at the corresponding frequencies (as in Peelle et al., 2013; Mai et al., 2016), but also for more complex acoustic (spectrogram) and higher-level linguistic (phonetics and phonemic) features.

4.2 | Distinct roles of delta- and theta-band neural tracking for neural sharpening and predictive coding

Cumulative evidence showed that speech tracking at delta and theta bands may play distinctive roles during processing of speech with various degrees of acoustic clarity or different linguistic contents (Ding

et al., 2014; Etard & Reichenbach, 2019; Mai et al., 2016; Molinaro & Lizarazu, 2018) as well as different aspects of predictive coding of contextual information (Donhauser & Baillet, 2020). It is therefore sensible to study speech tracking at these two frequency bands separately.

Our results showed that delta-band tracking was greater for pseudo-words than real-words, especially during phonetic encoding (Figures 4 and 5). While tracking for pseudo-words were also greater than time-reversed speech, no difference between real-words and time-reversed speech was found. Therefore, delta-band tracking may reflect predictive coding, such that tracking of utterances with more expected semantic content (i.e., real-words) was suppressed, while utterances with more unexpected semantic content (i.e., pseudo-words) was neurally tracked to represent the discrepancies between the heard speech and the predicted speech (i.e., prediction errors; Sohoglu & Davis, 2020). It might be argued, however, that participants should already realise the lack of valid words during an ongoing pseudo-word utterance. As a result, similar to 'Jabberwocky' stimuli (Kaufeld et al., 2020; Matchin et al., 2017, 2019; Pallier et al., 2011), no predictive effects of lexical semantics should be anticipated. However, it should be noted that our stimuli were not exactly 'Jabberwocky' stimuli. Jabberwocky syllables do not often have explicit meanings and/or potentials to form a valid word. For example, in an English Jabberwocky *sawl pand* (Matchin et al., 2017, 2019), *sawl* cannot form any valid word with another syllable. In contrast, our pseudo-words consisted of Mandarin morphologically valid syllables which are *all* commonly used morphemes that by themselves have certain semantic meanings with each having potentials to form a valid word with another syllable/morpheme. For example, in a pseudo-word '基米', syllable '基' (meaning *base*) has the potential to be followed by another syllable '本' (meaning *origin/root*) to form a valid (disyllabic) word '基本' (meaning *fundamental*). Hence, we argue there were involuntary/automatic priming effects (Deacon et al., 1999; Neely, 1977; Neely & Kahan, 2001) of a syllable to predict the one following it, despite no explicit manipulations of prior prediction. While probably not all syllables have such priming effects because of participants realising the lack of valid words during an ongoing pseudo-word utterance, greater semantic unexpectedness/surprisal *on average* is anticipated compared to real-words, which can explain the greater delta-band tracking observed here.

It is also noted that delta-band tracking of time-reversed speech was low and not significantly above chance for any feature (see Figure 3a). Theoretically, upcoming signals during listening to time-reversed speech are always highly unpredictable. The low tracking accuracies thus reflect poor encoding of the heard time-reversed speech under the predictive coding scheme. It means that sharpening may coexist along with predictive coding, such that features of time-reversed speech were poorly tracked due to its lack of linguistic content. Nonetheless, sharpening would always need to be in place as a prerequisite so that predictive coding can be implemented (Friston, 2005, 2012; Summerfield & De Lange, 2014).

Theta-band tracking, on the other hand, was greater for real-words and pseudo-words than time-reversed speech (Figure 4). Analyses at individual time lags showed that such effects were significant during encoding of all three features (Figure 6). This indicates that theta-band tracking contributes to neural sharpening possibly due to valid phonological and/or morphological contents in real-words and pseudo-words. These results echo the previous finding showing greater sharpening effect for neural tracking at theta than at delta band (Broderick et al., 2019). However, we did not find significant sharpening effects of semantics in the first instance (Figures 4 and 6a). As discussed above, all morphologically valid syllables should have word-level semantic priming effects to predict the next syllable to form a valid disyllabic word, so sharpening should be anticipated. We suspected that the lack of sharpening might be because only the first syllable within each disyllabic word had the priming effect so that sharpening only took effect on the second syllable. Therefore, measurements over the entire stimulus period may have smeared the sharpening effects. We thus separated theta-band tracking during the periods of the first and second syllables of disyllabic words and confirmed our suspicion. The theta-band tracking was significantly greater for real-words than pseudo-words at the second, but not the first, syllables of disyllabic words within the time lags of 200 ms (Figure 6b). There may also be further concerns that, in our real-words stimuli, words were not contextually related so that semantic predictability would not take effects. However, we argue that, because every real-word utterance had the same number syllables with a fixed syntactic structure (Subject + Verb + Adjective + [particle] + Object), participants should know the word categories at fixed positions (e.g., the second word in an utterance was always a verb). This therefore provided participants with further prior knowledge to enhance the semantic priming leading to potential sharpening effects (e.g., when participants heard the first syllable of the second word, they would be able to predict the second syllable based on the prior knowledge that a verb was to be formed). In sum, these results thus provide evidence for the role of theta-band tracking for neural sharpening during both phonological/morphological and word-level semantic processing.

While we found distinct roles of delta- and theta-band tracking for sharpening and predictive coding, questions may arise in terms of how such findings reconcile with previous studies with different findings. For example, Etard and Reichenbach (2019) revealed that delta- and theta-band tracking are respectively enhanced by the comprehensibility and acoustic clarity, indicating that, although their roles differ between various levels of speech processing, both are involved in neural sharpening. This is consistent with a large body of findings showing richer linguistic content related to greater delta-theta neural speech tracking (Coopmans et al., 2022; Gross et al., 2013; Keitel et al., 2018; Peelle et al., 2013; Tezcan et al., 2023). Another piece of evidence provided by Donhauser and Baillet (2020) showed that both delta and theta-band tracking play a role in predictive coding of contextual information. We therefore suggest that different observations may heavily depend on the experiment manipulations and speech features (e.g., acoustic and contextual manipulations/features) involved. We argue that the current findings would not preclude involvement of speech tracking at a certain frequency band for sharpening or predictive coding (e.g., delta-band tracking for sharpening and theta-band tracking for predictive coding), but instead stress that tracking at a

certain band may play a *greater* role than the other for sharpening/predictive coding.

It is also noteworthy that there is a possibility that the current observations may simply be due to attention effects. For example, greater delta-band tracking might be explained by greater listening efforts in difficult listening situations (Ding et al., 2014). Also, greater attention to utterances with richer linguistic content could lead to greater theta-band tracking. We tested this possibility by measuring the changes in parieto-occipital alpha-band power relative to pre-stimulus periods. We found the greatest negative change for time-reversed speech (Figure 5a), indicating that participants paid greatest attention to time-reversed speech, plausibly due to the greatest task difficulty. Therefore, the neural sharpening associated with richer phonological/morphological contents cannot be explained by greater attention to forward speech. Also, changes in alpha power did not differ between real-words and pseudo-words and the real-words versus pseudo-words difference in alpha power change was not correlated with the real-words versus pseudo-words difference in delta-band tracking (Figure 5b). Also, behavioural results (accuracies and reaction times) did not differ between real-words and pseudo-words (see Section 3.1). Hence, *no* evidence is shown that the difference in delta- and theta-band tracking between real-words and pseudo-words were the results of greater attention or listening effort. Although lack of alpha-band or behavioural effects might not provide strong enough evidence against the attention-based explanation, it is noteworthy that our results showed *opposite* effects of delta- and theta-band tracking. If attention is the main deterministic factor of neural tracking, we should anticipate similar effects of delta- and theta-band tracking. Alternatively, attention/listening effort might *partly* explain the results. For example, if attention is to explain greater delta-band tracking for pseudo- than real-words, greater theta-band tracking for real-words should be due to sharpening (greater expectedness); if attention is to explain greater theta-band tracking for real-words, greater delta-band tracking for pseudo-words should be due to prediction errors (greater unexpectedness). While we cannot totally exclude the role of attention/listening effort, our data has not provided evidence to support such role.

A further proposal to interpret the current findings may be the models of neural oscillations, for example, phase-aligning/phase-resetting of oscillations at different linguistic features (Benítez-Burraco & Murphy, 2019; Martin, 2020; Meyer, 2018; Peelle et al., 2013; Zoefel et al., 2018). While mTRF used here would reflect the mixed effects of oscillatory phase alignment/resetting and evoked responses to these features (Crosse et al., 2016), neural oscillations have been argued to play a major role in continuous speech perception (Meyer, 2018; Peelle et al., 2013; Zoefel et al., 2018). Indeed, the oscillation model is shown to better predict auditory cortical entrainment, for example, entrainment to music, compared to evoked responses (Doelling et al., 2019). During speech perception, low-frequency (delta and theta) oscillations are suggested to take a critical role by aligning their excitatory phases to important acoustic and linguistic features (Martin, 2020; Meyer, 2018; Zoefel et al., 2018). To interpret our findings based on this argument, delta-band oscillations may tend to be phase-aligned with more *unexpected* stimuli (pseudo-words) reflecting predictive coding while theta-band oscillations may be phase-aligned with more *expected* stimuli (real-words) reflecting neural sharpening. Indeed, previous work has proposed the models of predictive coding through oscillatory phase-alignment during neural processing of external stimuli (Arnal et al., 2015; Arnal & Giraud, 2012), including the processing of auditory speech in particular (Hovsepyan et al., 2020). Future work may look into how such models are in coordination with predictive coding and sharpening during neural processing of multi-level speech features and disentangle the effects between oscillations and evoked responses.

4.3 | Rapid computations of sharpening and prediction errors

Analyses on reconstruction accuracies across individual time lags showed that both neural sharpening and prediction errors started to appear at very early processing stages within 100 ms after feature onsets. For sharpening, this early processing took place during both phonological/morphological (theta-band tracking for real-words and pseudo-words vs. time-reversed speech during encoding of all three features; lower panels of Figure 6a) and semantic sharpening (real-words vs. pseudo-words when encoding all features combined; Figure 6b). For predictive coding, this *started* during phonetic encoding (delta-band tracking for pseudo-words vs. real-words) as opposed to mid-stage (~200–250 ms) during spectrogram encoding and late-stage (>400 ms) during phonemic encoding (upper panels of Figure 6a). Significant interactions between frequency bands (delta vs. theta) and stimulus types (real-words vs. pseudo-words) also started at this early stage during phonetic and phonemic encoding (Figure 6a, purple lines). Our results are thus consistent with Broderick et al. (2019) and Sohoglu and Davis (2020) which respectively show occurrence of sharpening and prediction errors at such an early processing stage. This also indicates such rapid computations can occur at higher-than-acoustic linguistic (phonetic) levels. Indeed, previous studies have shown that significant neural tracking of phonetic features occurs within 100 ms post-feature (Di Liberto et al., 2015; Teoh et al., 2022). In our case, prediction is updated in real-time *prior to* the speech inputs to be tracked, for example, the second syllable within a disyllabic word could be predicted just before the end of the first syllable, so that sharpening and prediction errors are formed through rapid computations at early stages.

Furthermore, we used these patterns of tracking accuracies across individual time lags to fit with the hierarchical Predictive Coding framework (Friston, 2005, 2012) in which neural information transferred between prediction errors of lower-level linguistic/acoustic features and sharpened signals of higher-level linguistic features (feedforward and feedback transfers, Figure 7a). According to the framework, prediction errors are feedforwarded from the lower hierarchical levels to the 'prediction units' at the higher levels to generate updated predictions with the sharpened signals; the feedback signals then transfer the sharpened signals back to the lower-level 'prediction error units' so that prediction errors are also updated (Friston, 2005, 2012). We showed that net feedforward transfers

occurred within 100 ms delays followed by feedback transfers occurring at ~150–200 ms delays (Figure 7b). These early feedforward and feedback flows thus indicate that, once prediction errors and sharpened signals are formed, they are swiftly utilised for information transfers within the framework. We interpret that such rapid computations and swift transfers may contribute to the recurrent updates of predictions and prediction errors over time (Friston, 2005, 2012) to support spoken language processing. Importantly, this hierarchical mechanism may also support long-term perceptual learning of speech (Sohoglu & Davis, 2016).

4.4 | Possible caveats and concerns

There are also several potential caveats that may need further attention for future work. First, differences between the real-words and pseudo-words here reflect the effects of word-level semantics. It is not clear how delta- and theta-band tracking play different roles in sharpening and predictive coding during processing of context-level semantics (Broderick et al., 2018, 2019; Broderick & Lalor, 2020). It is noteworthy that disyllabic words used in the real-word utterances were fitted in a syntactically valid structure, but they were not contextually related to each other (see Section 2.2). This may have caused possible unexpected contextual semantic effects resulting prediction errors in real-words. However, our results did not find such effects. It is possible that adequately 'surprising' rather than unrelated contextual information (Broderick & Lalor, 2020) are needed to result in such effects. Indeed, a recent study (Slaats et al., 2023) has provided evidence showing that delta- and theta-band neural signals differentially track lexical features depending on the existence of contextual information. Future work would be needed to use more contextual stimuli to separate the effects of word-level and contextual semantics.

The second is the naturalness of the stimuli. Neural tracking of speech has been widely researched using stimuli of long continuous speech, like audiobooks (e.g., Broderick & Lalor, 2020; Broderick et al., 2018, 2019; Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Etard & Reichenbach, 2019). This is more ecological and naturalistic with richer phonological and semantic variations than using short utterances as used here. Future work may thus combine the use of more naturalistic stimuli with more controlled traditional paradigms/stimuli to assess the roles of speech tracking at different frequency bands.

The third is the potential influence of tasks on the results. For example, a recent study (Ten Oever et al., 2022) examined how neural tracking of speech are modulated by different tasks. It was shown that delta-band tracking was greater when participants completed tasks that require attention to phrasal information compared to those that require attention to word and syllable information (Ten Oever et al., 2022). This thus indicate that attention to speech information at different linguistic levels or time scales could result in different strengths of neural tracking. In this current study, participants completed a sound/syllable-matching task hence syllable information was particularly attended to. It is not clear whether instructing participants to attend to different levels of linguistic information could modulate the current observed effects, which future work may need to study further.

A final critical concern may be the duration and number of training stimuli. Duration of each target period is 2 s (see Section 2.5.1). The average number of trials for training is ~48 (see Section 2.5.2). While it is unclear whether such duration and number were adequate to obtain robust neural tracking, we conducted additional analyses to see how many trials are needed to obtain stable patterns (how accuracies differ between stimulus types) consistent with the current findings (Section 2.5.2). We found that the patterns are highly consistent even when the number of training trials is as low as 60% of the total number. Same statistical results as the current findings can be obtained with 80% of the total number (see Figure S1 in the Supplementary Materials). Furthermore, the reconstruction accuracies were significantly above chance ($p < .0001$; see Section 3.2.1 and Figure 3b) and the values (>0.05 for delta-band tracking of pseudo-words and ~0.1 for theta-band tracking of real-words and pseudo-words) are comparable to previous reports that used lengthy audiobooks as stimuli (e.g., Di Liberto et al., 2015). These all support the validity of our results. However, this could still be surprising because the total duration of stimuli for training for each stimulus type is only ~96 s (~48 training trials, each 2 s long) compared to the suggested length of 10 minutes at minimum to obtain robust tracking (Di Liberto & Lalor, 2017). We argue that this may be because, first, the short utterance duration and nature of the task (a sound-matching task) ensured participants' attention to the stimuli compared to when a long and tedious audiobook is used. Second, the characteristics of Mandarin speech may result in particularly strong cortical tracking. As mentioned, all stimuli were produced at ~4 Hz syllable rate with all syllables having a similar duration of 250 ms. Despite this quasi-isochrony, they (the forward speech) still sounded relatively natural according to participants' feedback, arguably due to the syllable-timed nature of Mandarin (Mok, 2009). Previous studies using short sentences with similar syllable rhythms in Mandarin showed sharp and concentrated peaks of cortical tracking at the syllable (4 Hz) and word rates (2 Hz), respectively (Ding et al., 2016, 2018). We have also used the current dataset to show similar drastic concentrations of tracking at these frequencies (see Mai et al., 2016). Such phenomenon was not observed (i.e., more dispersed tracking values across delta-theta bands) when naturally produced stimuli in English, which is a stress-timed language, are used (e.g., Peelle et al., 2013). Importantly, mean cycles of the current delta (1.5–3 Hz) and theta (3–6 Hz) bands are exactly 2 and 4 Hz, respectively. Therefore, particularly strong cortical tracking at these two frequency bands may have made significant contributions to robust results despite the short total stimulus duration. Nonetheless, we suggest that longer stimulus duration and larger number of trials should consolidate our current findings.

4.5 | Summary

To the best of our knowledge, the current study is the first to show that delta- and theta-band tracking play possible distinctive roles for neural sharpening and predictive coding of multi-level speech features

during spoken language processing. Importantly, by applying the partialling procedures in multivariate temporal response functions, we investigated the *unique* effects for neural tracking of individual linguistic features (acoustics, phonetics and phonemes). This enabled us to study the brain processing of these linguistic features separately and examine how information representing sharpening and prediction errors were transferred across linguistic levels. We also investigated the time course of neural tracking after feature onsets to gain information as to when significant effects took place. Based on these, we provided novel insights into how sharpening and predictive coding may be operated across both times and linguistic levels during speech perception, which, as far as we know, have not been uncovered by previous research.

Particularly, we showed that delta-band tracking is involved with predictive coding driven by unexpected/unpredicted word-level semantic content, while theta-band tracking reflects neural sharpening driven by greater expectations of phonological/morphological and semantic content. Furthermore, we did not find evidence that these effects can be solitarily explained by attention or listening effort. We also showed that these effects started at early processing stages within 100 ms after feature onsets (sharpening during encoding of acoustics, phonetics and phonemes, or combined encoding of these features, and prediction errors during phonetic encoding), arguing for rapid neural computations of sharpening and prediction errors. Finally, we illustrate that our findings fit with the hierarchical Predictive Coding framework by showcasing the swift feedforward followed by feedback information transfers between prediction errors and sharpened signals across linguistic levels. The rapid computations together with the swift information transfers in the hierarchical framework may thus contribute to the recurrent updates of predictions and prediction errors over time to support spoken language processing. Taken together, this study revealed neural sharpening and predictive coding through neural tracking of continuous speech at different brain frequencies across times and linguistic levels. We suggest these findings contribute to the existing knowledge of predictive coding theory for spoken language processing.

ACKNOWLEDGEMENTS

The experimental work was supported by the Research Grants Council of Hong Kong (Grant No. 455911). Guangting Mai is jointly supported by NIHR Nottingham Biomedical Research Centre and UCL Language and Cognition. We thank Dr Aaron Nidiffer for providing the Matlab script that was used to perform the feature partialling procedure in combination of mTRF toolbox. We thank Prof Yun Mai (Institute of Linguistics, Chinese Academy of Social Sciences) to give critical advice on annotations of phonemic categories and classifications of phonetic features.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Original EEG data that support the findings of this study are available from the corresponding author upon reasonable request. Complete analyses codes can be found at https://github.com/guangtingmai/hbm26503/

ORCID

Guangting Mai https://orcid.org/0000-0002-5618-7420

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. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372.

Arai, T., Pavel, M., Hermansky, H., & Avendano, C. (1999). Syllable intelligibility for temporally filtered LPC cepstral trajectories. *The Journal of the Acoustical Society of America*, *105*(5), 2783–2791.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398.

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-beta coupled oscillations underlie temporal prediction accuracy. *Cerebral Cortex*, *25*(9), 3077–3085.

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

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Benítez-Burraco, A., & Murphy, E. (2019). Why brain oscillations are improving our understanding of language. *Frontiers in Behavioral Neuroscience*, *13*, 190.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*(11), e1002577.

Broderick, M. P., & Lalor, E. C. (2020). Co-existence of prediction and error signals in electrophysiological responses to natural speech. *bioRxiv*. https://doi.org/10.1101/2020.11.20.391227

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience*, *39*(38), 7564–7575.

Broderick, M. P., Anderson, A. J., di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Coopmans, C. W., de Hoop, H., Hagoort, P., & Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiology of Language*, *3*(3), 386–412.

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

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in Neuroscience*, *15*, 1350.

Daube, C., Ince, R. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924–1937.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

Deacon, D., Uhm, T. J., Ritter, W., Hewitt, S., & Dynowska, A. (1999). The lifetime of automatic semantic priming effects may exceed two seconds. *Cognitive Brain Research*, *7*(4), 465–472.

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

di Liberto, G. M., O'sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

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., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.

Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., & Zhang, J. (2018). Attention is required for knowledge-based sequential grouping: Insights from the integration of syllables into words. *Journal of Neuroscience*, *38*(5), 1178–1188.

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 10113–10121.

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393.

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University Press.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, *39*(29), 5750–5759.

Fraschini, M., & Hillebrand, A. (2017). Phase transfer entropy in Matlab. Figshare. *Dataset*. https://doi.org/10.6084/m9.figshare.3847086.v12

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology*, *83*(2), 248–252.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138.

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 Biology*, *11*(12), e1001752.

Hillebrand, A., Tewarie, P., Van Dellen, E., Yu, M., Carbo, E. W., Douw, L., Gouw, A. A., van Straaten, E. C. W., & Stam, C. J. (2016). Direction of information flow in large-scale resting-state networks is frequency-dependent. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(14), 3867–3872.

Hovsepyan, S., Olasagasti, I., & Giraud, A. L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*(1), 3117.

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience*, *40*(49), 9467–9475.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3), e2004473.

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., & Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *Hearing Research*, *380*, 1–9.

Lobier, M., Siebenhühner, F., Palva, S., & Palva, J. M. (2014). Phase transfer entropy: A novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *NeuroImage*, *85*, 853–872.

Londei, A., D'Ausilio, A., Basso, D., Sestieri, C., Gratta, C. D., Romani, G. L., & Belardinelli, M. O. (2010). Sensory-motor brain network connectivity for speech comprehension. *Human Brain Mapping*, *31*(4), 567–580.

Mai, G. (2014). Relative importance of AM and FM cues for speech comprehension: Effects of speaking rate and their implications for neurophysiological processing of speech. *Proceedings of Interspeech*, *2014*, 2585–2589. https://doi.org/10.21437/Interspeech.2014-554

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

Majdak, P., Hollomey, C., & Baumgartner, R. (2022). AMT 1.X: A toolbox for reproducible research in auditory modeling. *Acta Acustica*, *6*, 19.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427.

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, *40*(2), 663–678.

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621.

Mok, P. (2009). On the syllable-timing of Cantonese and Beijing Mandarin. *Chinese Journal of Phonetics*, *2*, 148–154.

Molinaro, N., & Lizarazu, M. (2018). Delta (but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Neely, J. H., & Kahan, T. A. (2001). Is semantic activation automatic? A critical re-evaluation. In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 69–93). American Psychological Association.

O'Sullivan, A. E., Lim, C. Y., & Lalor, E. C. (2019). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, *50*(8), 3282–3295.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Otnes, R., & Enochson. (1972). *Digital time series analysis*. John Wiley & Sons.

Pallier, C., Devauchelle, A. D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2522–2527.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1), e1001251.

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

Pereda, E., Quiroga, R. Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, *77*(1–2), 1–37.

Reichenbach, C. S., Braiman, C., Schiff, N. D., Hudspeth, A. J., & Reichenbach, T. (2016). The auditory-brainstem response to continuous, non-repetitive speech is modulated by the speech envelope and reflects speech processing. *Frontiers in Computational Neuroscience*, *10*, 47.

Santoro, R., Moerel, M., de Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology*, *10*(1), e1003412.

Saur, D., Schelter, B., Schnell, S., Kratochvil, D., Küpper, H., Kellmeyer, P., Kummerer, D., Kloppel, S., Glauche, V., Lange, R., Mander, W., Feess, D., Timmer, J., & Weiller, C. (2010). Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. *NeuroImage*, *49*(4), 3187–3197.

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

Slaats, S., Weissbart, H., Schoffelen, J. M., Meyer, A. S., & Martin, A. E. (2023). Delta-band neural responses to individual words are modulated by sentence processing. *Journal of Neuroscience*, *43*(26), 4867–4883.

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756.

Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *eLife*, *9*, e58077.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Ten Oever, S., Carta, S., Kaufeld, G., & Martin, A. E. (2022). Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. *eLife*, *11*, e77468.

Teoh, E. S., Ahmed, F., & Lalor, E. C. (2022). Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. *Journal of Neuroscience*, *42*(4), 682–691.

Tezcan, F., Weissbart, H., & Martin, A. E. (2023). A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. *eLife*, *12*, e82386.

van Dinteren, R., Arns, M., Jongsma, M. L., & Kessels, R. P. (2014). P300 development across the lifespan: A systematic review and meta-analysis. *PLoS One*, *9*(2), e87347.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wilmer, A., de Lussanet, M., & Lappe, M. (2012). Time-delayed mutual information of the phase as a measure of functional connectivity. *PLoS One*, *7*(9), e44633.

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, *32*(7), 855–869.

Wöstmann, M., Lim, S. J., & Obleser, J. (2017). The human neural alpha response to speech is a proxy of attentional control. *Cerebral Cortex*, *27*(6), 3307–3317.

Xu, L., & Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: Implications for auditory prostheses. *Hearing Research*, *242*(1–2), 132–140.

Zoefel, B., ten Oever, S., & Sack, A. T. (2018). The involvement of endogenous neural oscillations in the processing of rhythmic input: More than a regular repetition of evoked neural responses. *Frontiers in Neuroscience*, *12*, 95.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mai, G., & Wang, W. S.-Y. (2023). Distinct roles of delta- and theta-band neural tracking for sharpening and predictive coding of multi-level speech features during spoken language processing. *Human Brain Mapping*, 1–24. https://doi.org/10.1002/hbm.26503