Cerebral Cortex, 2023, **33**, 11287–11299
https://doi.org/10.1093/cercor/bhad366
Advance access publication date 6 October 2023
Original Article

EEG-based assessment of temporal fine structure and envelope effect in mandarin syllable and tone perception

Guangjian Ni[1,2,3,*,†], Zihao Xu[1,2,†], Yanru Bai[1,2], Qi Zheng[1], Ran Zhao[1,2], Yubo Wu[1], Dong Ming[1,2,3]

[1]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072 China,
[2]Tianjin Key Laboratory of Brain Science and Neuroengineering, Tianjin 300072 China,
[3]Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration, Tianjin 300392 China

[*]Corresponding author: Guangjian Ni: Email: niguangjian@tju.edu.cn

[†]Guangjian Ni and Zihao Xu contributed equally.

In recent years, speech perception research has benefited from low-frequency rhythm entrainment tracking of the speech envelope. However, speech perception is still controversial regarding the role of speech envelope and temporal fine structure, especially in Mandarin. This study aimed to discuss the dependence of Mandarin syllables and tones perception on the speech envelope and the temporal fine structure. We recorded the electroencephalogram (EEG) of the subjects under three acoustic conditions using the sound chimerism analysis, including (i) the original speech, (ii) the speech envelope and the sinusoidal modulation, and (iii) the fine structure of time and the modulation of the non-speech (white noise) sound envelope. We found that syllable perception mainly depended on the speech envelope, while tone perception depended on the temporal fine structure. The delta bands were prominent, and the parietal and prefrontal lobes were the main activated brain areas, regardless of whether syllable or tone perception was involved. Finally, we decoded the spatiotemporal features of Mandarin perception from the microstate sequence. The spatiotemporal feature sequence of the EEG caused by speech material was found to be specific, suggesting a new perspective for the subsequent auditory brain-computer interface. These results provided a new scheme for the coding strategy of new hearing aids for native Mandarin speakers.

Highlights

- Tone perception depended on the temporal fine structure.
- Syllable perception mainly depended on the speech envelope.
- The time-frequency-spatial brain response to Mandarin words is specific.
- Mandarin perception strongly relates to the parietal and prefrontal cortex.

Key words: auditory processing; speech envelope; temporal fine structure; task-state microstates; electroencephalogram.

Introduction

Mandarin is one of the most widely used tonal languages, and tonal modulation conveys different word meanings. Mandarin has four different tones: high-level (T1), mid-rising (T2), low-dipping (T3), and high-falling (T4) (Chao 1965). Different tones represent different meanings within the same syllable. For example,/bai T1/,/bai T2/,/bai T3/, and/bai T4/, which mean "split," "white," "swing," and "defeat," respectively. In daily life and communication, the accurate perception of Mandarin tones plays an important role. The investigation of whether the primary acoustic carrier of the four tones is the speech envelope (ENV) or the temporal fine structure (TFS) could be the theoretical basis for the new strategy of hearing aids.

Smith and Delgutte (2002) constructed a set of sound stimuli called an auditory chimera. Each auditory had the ENV of one audio and the TFS of another. This technique provided a way to study the relative importance of the ENV and the TFS in speech comprehension and the perception of tones (Roy et al. 2015; Deroche et al. 2019). The temporal information of the speech signal could be divided into ENV and TFS based on the Hilbert transform. The ENV is defined as the amplitude profile of the speech signal, and the TFS is defined as the instantaneous phase information related to the harmonic resolution in the signal (Kong and Zeng 2006). The TFS is usually called the "carrier," and the ENV is generally called an amplitude modulator applied to the carrier. In the auditory system, the ENV cue shows the fluctuation of the short-term discharge rate of auditory neurons, while the TFS shows the synchronization of the nerve spike and the carrier-specific phase (Rose et al. 1967; Joris 1992). At present, cochlear implants provide envelope information through relatively few effective channels. Non-tone users acquired good speech understanding in a quiet environment (Wilson et al. 1991). The speech perception performance of Mandarin users was lower than that of English and other non-tone languages users because the coding strategy of cochlear implants was mostly Continuous Interleaved Sampling, which mainly used the envelope information of speech signals and sinusoidal carrier modulation as speech stimuli. The current improved fine-structure coding strategy (FSP) with four channels increased speech and music perception of non-tone languages users in a noisy environment and tone perception of tonal languages users (Roy et al. 2015; Meng et al. 2016; Vandali et al. 2019).

Received: July 22, 2023. **Revised:** September 13, 2023. **Accepted:** September 15, 2023

Previous psychological studies have shown that the ENV is sufficient for speech perception in quiet, whereas the TFS is essential for perceiving speech and tones in noise (Smith et al. 2002; Xu and Pfingst 2003; Hopkins et al. 2008; Hopkins and Moore 2009; Apoux and Healy 2013). Functional magnetic resonance image (fMRI) and positron emission computed tomography (PET) experiments, which used tone as a stimulus, showed that left hemisphere activation was more pronounced in tonal languages speakers compared to native English speakers (Klein et al. 2001; Gandour et al. 2002; Wong et al. 2004), suggesting that hemisphere lateralization depended on language function and language experience rather than acoustic properties. Since the temporal resolution of the hemodynamic response measured by fMRI or PET ranges from a few seconds to tens of seconds (Kim et al. 1997; Amaro and Barker 2006), the observations from these neuroimaging studies might represent the brain activity of temporal aggregation, including the attentional stage of auditory processing. However, auditory speech processing might occur within a short time (within 400 ms) after the onset of stimuli. Therefore, earlier fMRI and PET studies did not show the brain changes within 400 ms after sound stimuli. Later, an electroencephalogram (EEG) was introduced to study brain response changes in the early period after speech stimuli. Luo et al. (2006) showed that the early auditory processing of tones in the pre-attentional stage was biased to the right hemisphere based on mismatch negative (MMN). Speech can be decomposed into two parts (ENV and TFS) using a "filterbank" method. The conventional wisdom is that speech perception is largely dependent on ENV and that TFS improves speech perception in complex environments (Hopkins and Moore 2009; Apoux and Healy 2013; Prinsloo and Lalor 2022). Current neurophysiological studies have shown that TFS contributes to robust cortical activity in complex environments (Ding and Simon 2014). Indeed, speech signals after elimination of ENV information can still elicit cortical responses. From a neurophysiological perspective, it has been suggested that the TFS provides tonal cues that contribute to speech recovery of critical temporal information and significantly improves speech intelligibility, helping to form auditory objects for segregation and integration (Kong and Zeng 2006; Hopkins et al. 2008; Teng et al. 2019).

In contrast, the study of consonants showed that auditory processing was biased to the left hemisphere, which shows that tones and consonants depend on different acoustic features (Luo et al. 2006). Speech understanding requires extracting acoustic features from sound signals in real time and converting them into speech representations such as syllables, words, and sentences. Cortical activity in the delta and theta bands has been found to track speech rhythm (Ding and Simon 2014; Goswami 2019). However, whether cortical speech tracking involves high-level language processing is controversial. Previous studies using English stimuli materials showed that theta band cortical speech tracking encodes speech intelligibility, followed by acoustic aspects of signal light, while delta band speech tracking encodes more advanced speech perception (Etard and Reichenbach 2019). Ho et al. (2019) conducted an event-related potential (ERP) study on Mandarin, which showed that Mandarin recognition was gradual, and tone information was still speech information, but it would be recognized when tone information became available after the word appeared.

Presently, research on brain response in Mandarin focuses mainly on ERP and EEG rhythm and rarely examines the brain response of syllables and tones in terms of time and space. Microstates could reasonably be parsed into a series of stable time intervals in the sub-second range, reflecting a distinct conscious mental brain state. EEG microstate analysis has been seen as a practical approach to studying the neural signatures of many cognitive processes. For example, microstate dynamics have been associated with perceptual awareness (Britz et al. 2014), visual processing (Britz and Michel 2011), neuropsychiatric disorders (Lehmann et al. 2005; Kindler et al. 2011), and disorders of consciousness (Gui et al. 2020). Currently, most cognitive and disease researches are concerned with analyzing resting state microstates, which are generally divided into four categories. These four types of microstates could explain >80% of the brain's resting state (Khanna et al. 2015; Mishra et al. 2020). Gui et al. (2020) analyzed the task-state microstate of the speech paradigm and found specific assessment indicators to assess the unconscious state, conscious state, and normal people's microstate, which could be used to evaluate the brain state of vegetative people. This work provided theoretical support for the subsequent task state analysis of syllables and tones.

The current debate has mainly focused on which EEG rhythm, which brain area, and whether there are specific temporal–spatial features in tone perception. Previous studies showed the importance of TFS for tone perception and recognition and ENV for speech perception. Two hypotheses were proposed: (i) Tone recognition mainly depends on the TFS, and the brain response elicited by different tones has different time-frequency-spatial characteristics; (ii) Syllable recognition mainly depends on the ENV, and different syllables elicit different time-frequency-spatial brain response characteristics. In this paper, the ENV and TFS of speech were extracted using the auditory chimera, and the syllable-tone perception EEG experimental paradigm of Mandarin was designed to explore the time-frequency-spatial characteristics of the brain response of speech perception. The experimental paradigm was divided into three acoustic conditions: original speech, ENV, and TFS; phonetic materials were Pinyin: the four tones of "ba," "yao," and "yuan" making up 12 stimuli, as listed in Table 1. The above hypotheses were further verified by analyzing the time-frequency-space characteristics of syllable and tone EEG data of auditory evoked potentials (AEP), power spectral density (PSD), and microstate, as shown in Fig. 1.

Materials and methods

Participants

A total of 21 native Mandarin speakers participated in this experiment (23.5 ± 4 yr old; 9 female). The experiment was undertaken following the Declaration of Helsinki. All subjects had given their written informed consent to participate in the research, and the Ethics committee of Tianjin University approved the experimental procedure and experimental paradigm. Subjects reported no history of hearing impairment or neurological disorder.

Stimuli

The original language test materials consisted of three Mandarin monosyllables. The Pinyin (i.e. the phonemic writing system for Mandarin) of these syllables were "ba," "yao," and "yuan," and each syllable had four different tones: T1, T2, T3, and T4. The auditory materials consisted of 12 words narrated by Mandarin speakers. Each word was recorded using a Sennheiser e845S with a sampling frequency of 48,000 Hz. The sound was set to the same root mean square (RMS) and was perceived as the same intensity.

In our experiment, we used auditory chimeras (Smith et al. 2002) to decouple the envelope of the speech stimuli from the

Table 1. Speech words were used as stimuli.

Syllable	Lexical Tone	Meaning/Translation	Duration (ms)
Ba	T1	Eight	175
	T2	Draw	254
	T3	Target	289
	T4	Dad	217
Yao	T1	Goblin	247
	T2	Kiln	283
	T3	Bite	312
	T4	Drug	256
Yuan	T1	Injustice	246
	T2	Element	287
	T3	Far	308
	T4	Enmity	248

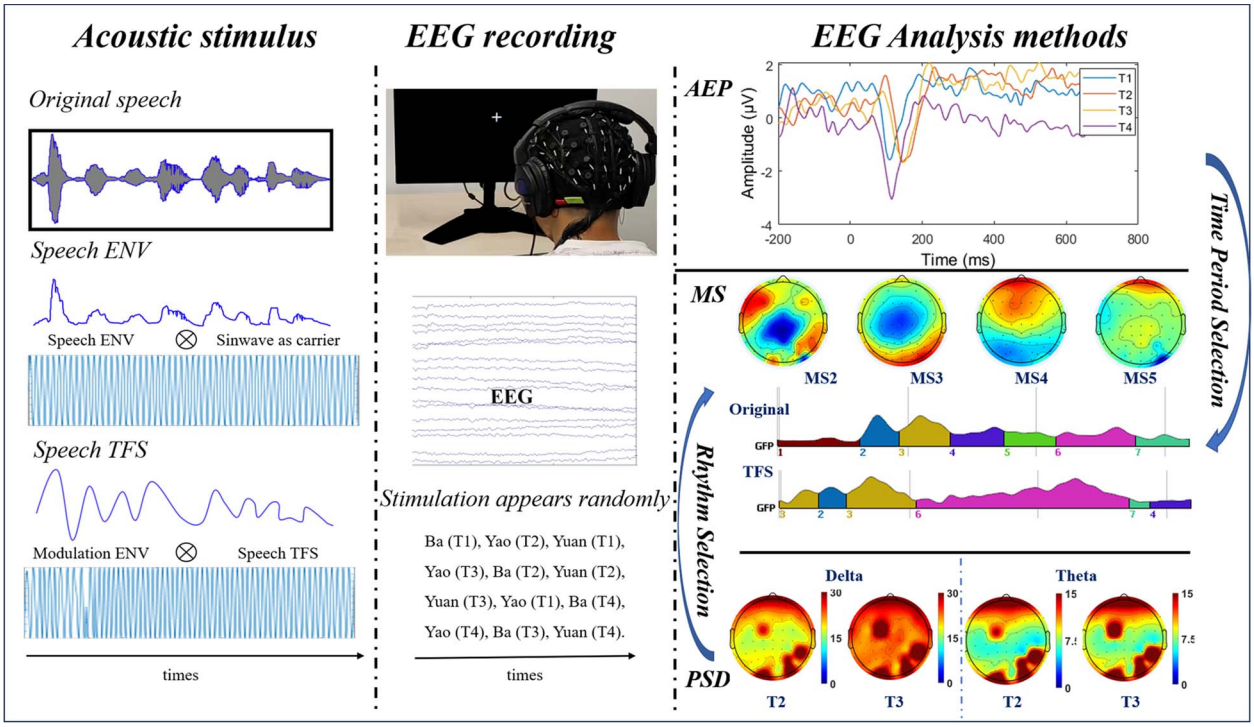

Fig. 1. An overview of the experiment and analysis. Subjects participated in 12 mandarin material perception experiments. The experiments included original speech, ENV, and TFS. The ENV and TFS were created with the summed output of a bank of 20 bandpass filters from 100 to 8,000 Hz. Behavioral data and EEG responses were obtained at the three acoustic conditions. AEP, PSD, and microstate sequence analysis were used to investigate the importance of ENV and TFS in mandarin perception, in which AEP provides the time domain reference range for task state microstate analysis, and PSD provides the EEG rhythm range.

fine structure. The original material was filtered by log-linear gammatones (20 filter banks) in the range of 100–8,000 Hz, and the output of each filter was transformed by the Hilbert transform to obtain its analytic signal-speech ENV (calculated as the magnitude of the analytic signal) and TFS (calculated as the cosine of the phase of the analytic signal). We needed to use the TFS in each frequency band and the non-speech sound (white noise) envelope in the corresponding frequency band to perform a convolution operation, resulting in a series of partial chimeras. The newly generated chimera retains the TFS of the original speech without introducing any other meaningful envelope information, which was convenient for us to explore whether the ENV or the TFS dominated speech perception. Finally, all frequency bands were added to produce the final speech stimuli material. MATLAB was used for all signal processing.

Experiment design of the auditory perception study

All tests were performed in a quiet room, and the stimuli were presented to the listeners monaurally through Sennheiser HD280 headphones. The original and processed speech (ENV and TFS) was presented at a fixed level of ∼65 dBA. The speech signal was transmitted through the main test computer's UR-22C sound card (ASIO). The experiment was divided into three sub-experiments using different stimuli materials: original speech, ENV, and TFS.

Each experiment started with a visual fixation cross in the center of a computer monitor for 1,000 ms. The EEG data of this period were taken as resting state. Each subject listened to 12 words of speech, each appearing 15 times for 180 speech stimuli. These 180 speech stimuli were played pseudo-randomly (to ensure that the same words were not played continuously). After 1,000 ms of each

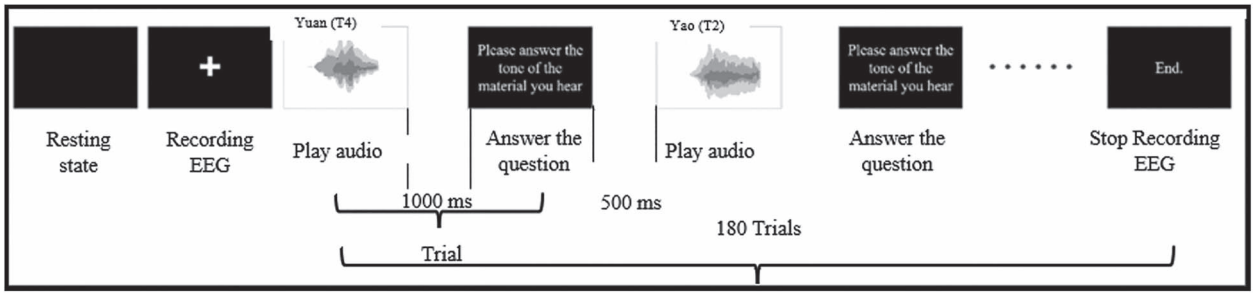

Fig. 2. Experimental paradigm diagram.

speech playback, the question "Please select the tone of the speech that you have just heard" appeared on the monitor, the number key corresponding to the answer should be pressed, and the key response was recorded. This setting was to avoid the interference of EEG artifacts caused by the execution of the movement. It also improved the quality of the EEG signals. The entire experimental procedure and the behavioral data record were written using Psychtoolbox. The experimental procedure is shown in Fig. 2. The total duration of Experiment 1 was ∼10 min. In Experiment 2 and Experiment 3, the speech stimuli material was replaced by the corresponding processing speech, and the subsequent experimental procedure was the same as in Experiment 1. This setting aimed to investigate the role of the ENV and the TFS in tone perception (especially on T2 and T3 perception) (Wang et al. 2013). The total duration of the experiment was ∼30 min, and the rest could be taken between the experiments according to the subjects' needs to ensure that the subjects were in good condition to participate. The complete experimental materials are listed in Table 1.

EEG recording and preprocessing

EEG acquisition

The EEG was recorded from 64 electrodes using NeuroScan SynAmps2 according to the international 10–20 system (EEG data of M1, M2, CB1, ad CB2 channels were not adopted in this work due to it was not needed as a reference, and positioning the cerebellum.). The EEG marker corresponds to the label of the stimuli when the speech stimuli start to play (e.g. "yao T1" was labeled "1."). The digital sampling rate was 1,000 Hz with an online 0.1–150 Hz frequency bandpass filter and 50 Hz notch filter, and all data were referenced to the average of all scalp channels. A total of 30 min of EEG data were obtained for each subject. All electrode impedances were maintained below 10 kΩ.

Data Preprocessing

For the AEP and PSD analysis, EEG data were preprocessed using MATLAB (version 2021) as follows: data were bandpass filtered (0.1–30 Hz) with a notch filter (50 Hz) firstly. Then, channels were semi-automatically inspected, and bad ones were interpolated. Next, data were re-referenced to the common average of signals from all EEG channels, and an independent component analysis (ICA) was performed to remove blinks, eye movements, and electromyography (EMG) signals. Finally, data were segmented to 1.2 s (200 ms before and 1,000 ms after the label) epochs and downsampled to 200 Hz.

For the brain state analysis, EEG data were preprocessed in the EEGLAB toolbox (version 2021) as follows: the electrodes were placed in the mastoid and cerebellum first, and data of the maintained 60 electrodes were bandpass filtered (0.2–40 Hz). Then, channels were semi-automatically inspected, and bad channels were interpolated before. Next, ICA was performed to remove blinks, eye movements, and EMG. The data were segmented into 800 ms epochs and manually removed bad epochs. Finally, data were re-referenced to the common average, and bandpass filtered again (1–20 Hz).

Brain microstates analysis

The microstate analysis included checking the scalp data of the brain according to a set of fixed brain topographic maps and quantifying the data according to the dominant period of these topographic maps. Microstate analysis belongs to the decomposition of data from the time–space perspective. It decomposed the multi-channel EEG time series into a relatively small group of spatial topographic maps or a time-varying linear combination of components. It achieved independence by eliminating the overlap of each group in time so that there was only one brain topographic map at a given time (Michel and Koenig 2018). Microstate analysis typically consists of a clustering step, where the scalp field data to be analyzed is submitted to a spatial clustering algorithm that identifies the component/microstate maps, and an assignment step, where the individual time points and conditions of the data are assigned to the best-fitting cluster, yielding the time-courses of the microstate model (Khanna et al. 2015; Michel and Koenig 2018; Gui et al. 2020; Mishra et al. 2020).

Different microstates represent different activations from underlying neural sources, so they could be argued to reflect different cognitive processes (Khanna et al. 2015). Microstate analysis could be used to study whether the time of some brain processes was different between factor levels, that is, whether their length, attack time, or effective latency were affected by the system of experimental operation. To this end, we focused on the following components:

N1: An enhanced negative wave recorded during selective attention to an auditory signal. The N1 was related to the physical parameters of the auditory stimuli, and its amplitude and latency directly reflected the human brain's perception and processing of information to the sensory input. Therefore, in our study, it was expected that the physical properties of different sound materials would respond differently ∼100 ms after the onset of stimuli. The main area of the N1 was in the common cortex on the side of the temporal and parietal cortex.

P300: Many studies have shown that the P300 was an ERP component related to cognitive functions such as attention, recognition, decision-making, and memory. The latency of the P300 was thought to reflect the time of sound processing by the cortical auditory system. Therefore, in our study, it was expected that stimuli samples of different syllables and tones would differ in the prefrontal and parietal lobes at ∼300 ms.

N1-P2 complex wave: The source of the N1-P2 complex wave was located in the upper part of the standard head temporal lobe, with a latency of ∼100–200 ms. Broadband or narrowband sound with different sound correlations could produce stable and obvious N1-P2 components. Therefore, N1-P2 could be used as a

brain topographical map to be observed for tone and syllable in time difference.

In this study, the number of subjects was 21 among the original, ENV, and TFS groups. Common templates were used to avoid systematic differences caused by differences in group-specific microstate templates between different groups. The improved T-AAHC clustering algorithm was applied to the brain topographic map clustering to obtain a common template. According to different research contents, the best template was clustered. The GEV, KL, and MeanCrit criteria were used to determine the optimal number of clusters, and the brain map was obtained by clustering in turn.

The obtained optimal microstate template was back-fitting to the EEG data of each subject, and the EEG map was converted into a microstate sequence. Temporal smoothing was then employed in the microstate sequence by changing the labels of small segments (<20 ms) to the next most likely microstate class until no microstate segment was smaller than 20 ms (Liu et al. 2020). By back-fitting the clustered microstates map, we calculated the global explained variance (GEV), duration, coverage, onset, offset, and microstate centroid. A detailed description of the microstate parameters can be found in the study by Khanna et al. (2015) and Michel and Koenig (2018). In the subsequent microstate analysis and result discussion, we mainly focused on the change of microstate parameters in the time range of 100–400 ms.

Statistics analysis

In the study, the significance level was 0.05. All the reported *P*-values were based on the non-parametric permutation method. For multiple comparisons, we reported the *P*-values corrected by the Bonferroni method. The paired t-test was used to compare the tone between groups in behavioral studies. The latency and amplitude of AEP syllables and tones were analyzed by ANOVA and paired t-test, and one-way ANOVA analyzed the latency and amplitude of stimuli. In the PSD analysis, the statistical method of EEG rhythm between different electrodes was paired t-test.

In the brain state analysis, the main effect of the group under each condition was examined using ANOVA for the parameters of microstates. A repeated-measures ANOVA was used to evaluate the group effect in each task condition, whereas group (three tasks: original control, ENV, and TFS) was the between-participants factor. For pairwise comparisons between the three groups, one-way ANOVA tests (Bonferroni corrected) were applied to all EEG metrics in each condition.

Results

Behavioral assessment

We first investigated the ability to perceive the tone of Mandarin under the stimuli of different speech chimeras. We used the time domain envelope information of sinusoidal carrier modulation speech analysis to synthesize speech envelope material and the TFS information of non-speech (white noise) envelope modulation to synthesize speech fine structure material. Three acoustic conditions, namely original speech, ENV, and TFS were used to characterize different tone perception levels. Statistical analysis was performed on the tone perception behavior data of 21 subjects who participated in the experiment, as shown in Fig. 3. Figure 3(a) shows the behavior result of the original speech. The most easily recognized tone was T4, with an accuracy rate of 95.67%; the accuracy rate of T3 was the lowest, only 76.84%, and most of the misjudgments were T4 (51.51% of misjudgments). Figure 3(b) shows the behavioral results of the subjects of speech envelope, from which we can see that the accuracy rate of T1 was up to 86.68%; the accuracy rate of T2 was the lowest, only 18.01%, and most of the T2 was misjudged as T1 (83.45% of the misjudgments). It can be seen from the results of the confusion matrix that most of the misjudgments of four tones were T1. The behavioral results showed that the envelope signal was not the main carrier of tonal information. Figure 3(c) shows the behavior results of the subjects with TFS, from which we can see that the accuracy rate of each tone recognition is >84%, and the tone information could be well perceived. However, it could be seen from the tone accuracy rate of misjudgment that the T2 and the T3 were relatively easy to confuse, which might be the reason that the spectrum of T2 and T3 were similar, and the physical properties of the sound itself ed to easy confusion (Luo et al. 2006; Li et al. 2021). The overall decoding accuracy of the three experiments is shown in Fig. 3(d). The accuracy of the original speech and the TFS was better than that of the speech envelope, and statistical differences existed between them. The behavioral results showed that the TFS was the primary carrier of the tone information, and the TFS was better than the original speech. This might be due to the reduction of redundant information features, which improves tone recognition accuracy.

Cortical tracking of syllables and tones

As the first step in investigating syllable and tone perception from EEG recordings, we quantified the response of different EEG components to syllable and tone perception under the original speech stimuli. In particular, we analyzed syllable and tone perception through the N1-P2 joint component of the EEG response and statistically analyzed the response of C4 and C3 electrodes (Luo et al. 2006). Based on previous anatomical results examining C3 and FC3 as auditory-related brain regions and the results of brain response lateralization in Mandarin, and in conjunction with the time-domain topographic maps (shown in Fig. 4) and power-spectrum statistics of the present study, it was concluded that the main observations in this study were the C3, C4, and FC3 channel brain responses(Klein et al. 2001; Kong and Zeng 2006; Zhang et al. 2012b; Yu et al. 2014; Ho et al. 2019).

We found that the most significant EEG component of AEP in response to syllables was the N1-P2 complex wave, and the EEG response caused by different syllables was different, consistent with previous research (Ho et al. 2019). We preprocessed the collected syllable EEG signals and the C4 electrode results, as shown in Fig. 5(a). It can be found that the latency of N1 was 170–220 ms with a magnitude of −1.5 to −3μv, and the latency of P2 was 250–330 ms with a magnitude range of 0.5–4μv. The statistical analysis showed that the AEP response had a statistical difference in P2 but not in N1. The statistically paired t-test results of P2 are shown in Fig. 5(c). The average latency and amplitude of syllable "ba" were 311.9 ms and 1.9 μv, "yao" 316.8 ms and 2.05 μv, and "yuan" 329.3 ms and 2.09 μv. The statistics of syllable "yao" and "yuan" was $t(20) = 2.732$, $P = 0.0128$, and "ba" and "yuan" was $t(20) = 3.241$, $P = 0.0041$. Therefore, the latency of P2 could be considered the characteristic of EEG for syllable identification.

Similarly, we preprocessed and averaged the AEP response of the tone and examined the temporal waveform of its brain response, as shown in Fig. 5(b). We concluded that similar to syllables, the complex wave of N1-P2 was also the main feature of different tones in the time domain. Compared with the AEP waveform of syllables, the latency of N1 caused by different tones was relatively concentrated, and the amplitude difference was smaller. We also conducted statistical analysis on the latency and amplitude of N1-P2 of the C4 electrode and found that the

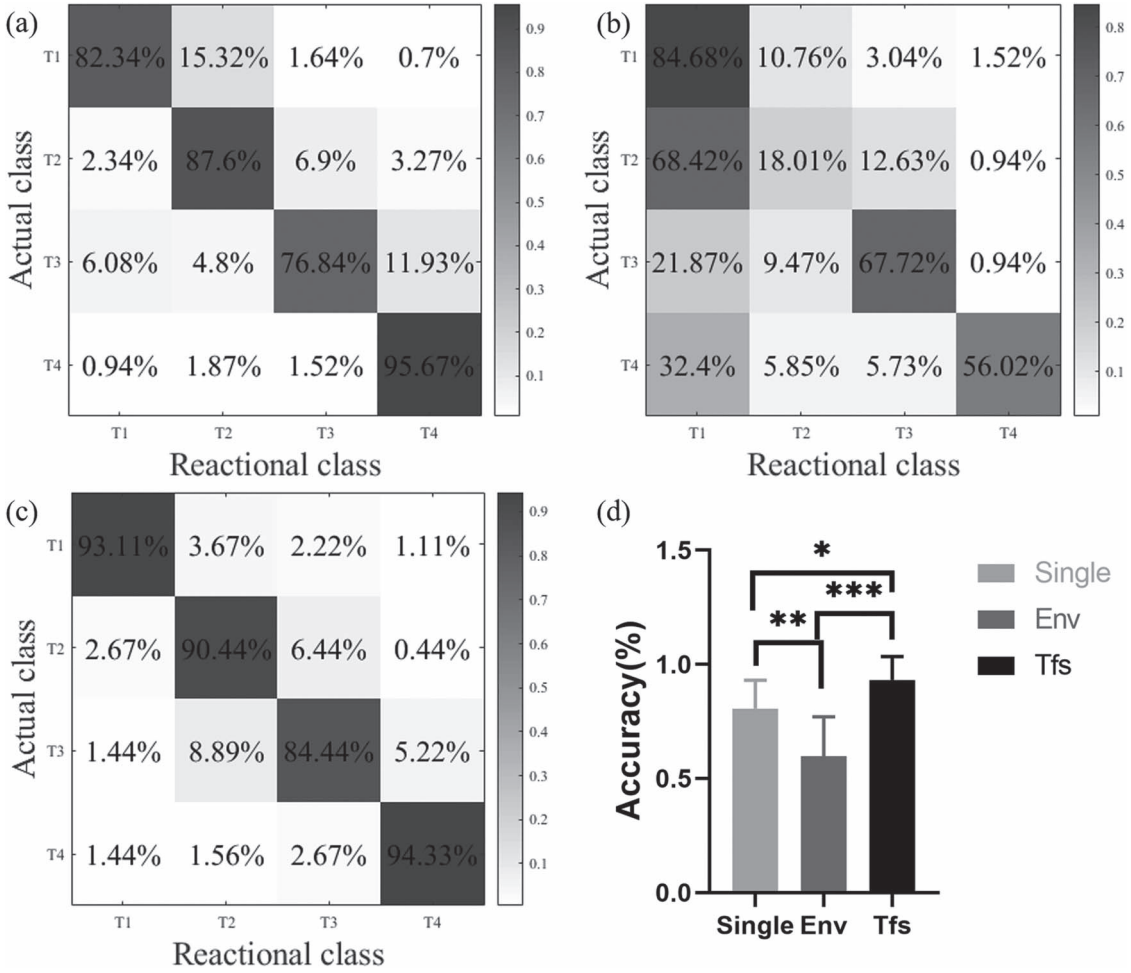

Fig. 3. Behavioral results. (a)–(c) Shows the behavioral results of the original speech, ENV, and TFS. (d) Shows the statistical results of the mean value of tone perception at three acoustic conditions. The thick dotted line indicates the second quartile and the thin dotted lines indicate the first and third quartiles. Statistical significance exists between the original speech and ENV and between TFS and ENV (paired t-test, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$).

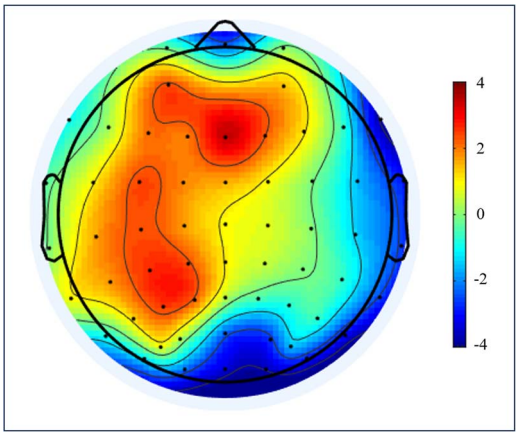

Fig. 4. Brain topography of speech stimulation material.

latency of P2 was statistically different, while the latency and amplitude of N1 and amplitude of P2 were not statistically different. The P2 latency of T1 was 311.2 ms, T2 327.3 ms, T3 326.0 ms, and T4 305.5 ms. We performed a one-way ANOVA for tones, $P < 0.0001$($F(2.661, 149.0) = 8.282$). A pairwise paired t-test for tone showed that $t(20) = 3.638$, $P = 0.0089$ for T1 and T2, $t(20) = 2.651$, $P = 0.0356$ for T1 and T3, $t(20) = 3.826$, $P = 0.0030$ for T2 and T4, and $t(20) = 3.790$, $P = 0.0063$ for T3 and T4. Therefore, time domain P2 latency could be used as a biomarker for tone identification. In the microstate analysis, we focused on analyzing the microstates of the brain response within 400 ms based on the EEG temporal results. Examine the statistical differences in the temporal and spatial distribution of the microstates of different stimuli.

In the time domain analysis, we investigated whether the brain response caused by twelve stimuli could find the temporal EEG characteristics. Therefore, after analyzing the EEG data of 12 speech stimuli samples, we found that C3, C4, and FC3 could be used as the time-domain characteristic of speech stimuli, as shown in Fig. 4. We performed one-way ANOVA and paired t-tests on the amplitude and latency of P2 in 12 stimuli of C4, as shown in Fig. 6. There was no statistical difference in the amplitude of P2, while the univariate analysis of variance of latency $P < 0.0001$ ($F(5.92, 1,06.6) = 13.72$) showed a statistically significant difference. This further shows that our experimental paradigm was reasonable and verified the first hypothesis mentioned above, i.e. that the brain response caused by different speech materials was specific. There were seven pairs with no statistical difference in the paired t-test, mainly because there was no statistical difference in the P2 latency of T2 and T3 of each syllable, which was consistent with the analysis of the behavioral results of the original speech,

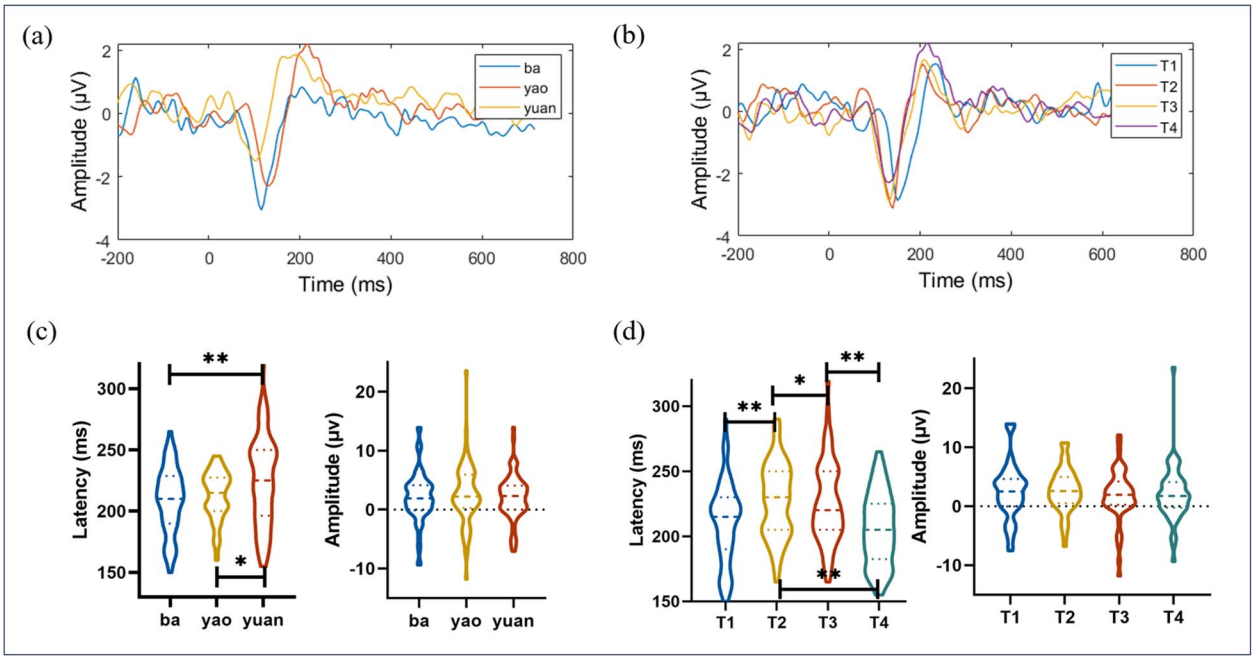

Fig. 5. Syllable and tone perception AEP results. (a) and (b) are the results of the syllable and tone perception AEP in the C4. The brain response caused by different speech materials was specific, especially N1-P2. (c) Shows the statistical results of the peak latency and amplitude of syllable perception P2. $P < 0.05$ for univariate ANOVA for the three-syllable peak latency. The paired t-test results are shown in the figure. The P2 amplitude was not statistically significant. (d) Shows the statistical results of sound perception's peak latency and P2 amplitude. $P < 0.001$ for univariate ANOVA for the three-syllable peak latency. The results of the paired t-test were shown in the figure, and the amplitude was also not statistically significant. (*$P < 0.05$, **$P < 0.01$).

that is, the T2 and T3 were relatively easy to confuse and difficult to distinguish.

PSD analysis of syllable and tone

We first analyzed the PSD of syllable and tone perception at the three acoustic conditions: original speech, ENV, and TFS. We used the PSD algorithm to analyze the EEG signal at 0.1–35 Hz, explored the EEG rhythm related to syllable and tone perception, and drew the brain topographical map, as shown in Fig. 7. Figure 7(a) shows that the syllable perception of the original speech was mainly delta (1–4 Hz) and theta (4–8 Hz) EEG rhythms, and the main brain areas were prefrontal, parietal, and temporal lobes. We statistically analyzed the electrodes in these three regions and found that the C3 and C4 could be used as significant features of syllable differentiation. The findings were consistent with previous studies (Luo et al. 2006). We performed a paired t-test on the PSD of different syllables in the C3 and C4, in which "ba" and "yao" were $t(20) = 10.330$, $P < 0.0001$ in the C4 and $t(20) = 4.752$, $P = 0.0003$ in the C3; the $P < 0.0001$ of "yao" and "yuan" in the C4 and the C3 ($t(20) = 14.389$ in the C4, $t(20) = 12.387$ in the C3). Figure 7(b) shows the result of the PSD analysis of the EEG signal of ENV. We found that the recognition of ENV syllables was mainly delta rhythm, and the main brain areas were the temporal and parietal lobes. Similarly, the paired t-test was carried out for C4 and C3. We found that the C4 "ba" and "yuan" was $t(20) = 2.859$, $P = 0.0371$, that of "yao" and "yuan" was $t(20) = 2.259$, $P = 0.0266$, and that of C3 was $P < 0.001$("ba" and "yuan" was $t(20) = 5.484$, "yao" and "yuan" was $t(20) = 6.892$). Figure 7(c) shows the PSD analysis of the TFS for syllable identification. The result showed that there was no statistical significance, that is, the syllable information cannot be recognized. Based on the above results, we could conclude that the main carrier of syllable information was the envelope of speech rather than the fine structure.

According to previous behavioral results, the most challenging part of tone recognition was T2 and T3 recognition (Wang et al. 2013). For this reason, we would examine the T2 and T3. The results are shown in Fig. 8, where the same analysis method was applied to T2 and T3 identification. By analyzing the original speech EEG data, we could conclude that tone recognition mainly depends on delta, theta, and alpha (8–12 Hz) EEG rhythms. The brain areas involved were the prefrontal and temporal lobes. We carried out a statistical analysis on the FC3, C3, and C4, and the results were as follows: (i) The delta, theta, and alpha EEG rhythms of the T2 and T3 of the original speech in FC3 had statistical significance; (ii) The T2 and T3 of the TFS had statistical significance in C4 and FC3; (iii) The ENV had no statistical significance in T2 and T3. It could be concluded from Fig. 8, which was also consistent with the behavioral conclusions, that the main carrier of tonal information was the TFS of the speech rather than the ENV.

Spatiotemporal analysis of syllables and tones—Microstate results

We used the best clustering of the microstate template to partition the EEG data from syllable and tone perception experiments of 21 subjects under three acoustic conditions. We independently applied the optimal microstate template to three conditions of syllables and tones, allowing us to identify potential topographic maps specific to a particular group.

We found the seven best equivalent topographic maps in the syllable perception experimental data, as shown in Fig. 9(a). From the AEP results, we could conclude that the main time of the syllable task state was between 100 and 400 ms so we would focus on the analysis of microstates in this period and microstates 2–4. From the microstate topography map, we know that microstates 2–4 mainly involve parietal and prefrontal lobes, consistent with previous studies on syllable perception brain response (Etard and Reichenbach 2019; Ho et al. 2019; Brohl and Kayser 2021; Prinsloo and Lalor 2022; Wei et al. 2022). For the main characteristic parameters of microstate, GEV, duration, and coverage, we performed

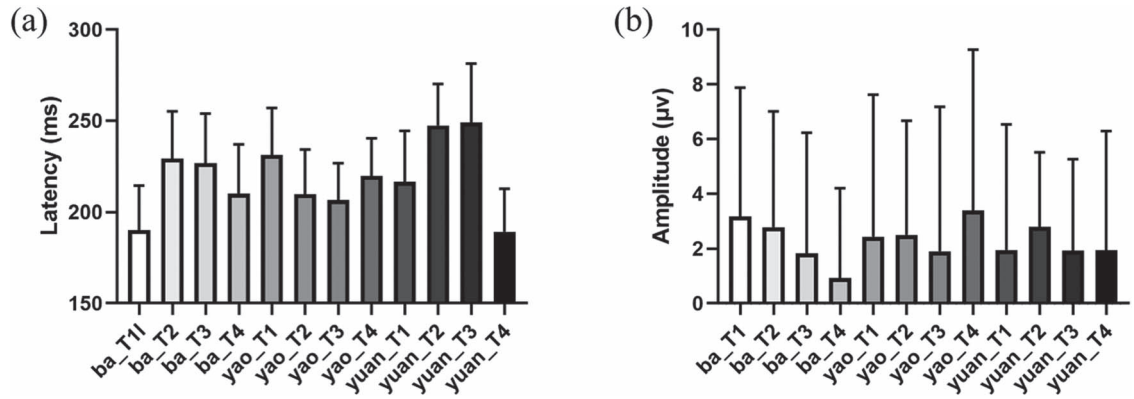

Fig. 6. P2 results of 12 speech materials. (a) the result of P2 amplitude latency. $P < 0.001$ for univariate ANOVA for the three-syllable peak latency. (b) The result of P2 amplitude. The univariate analysis of variance had no statistical difference.

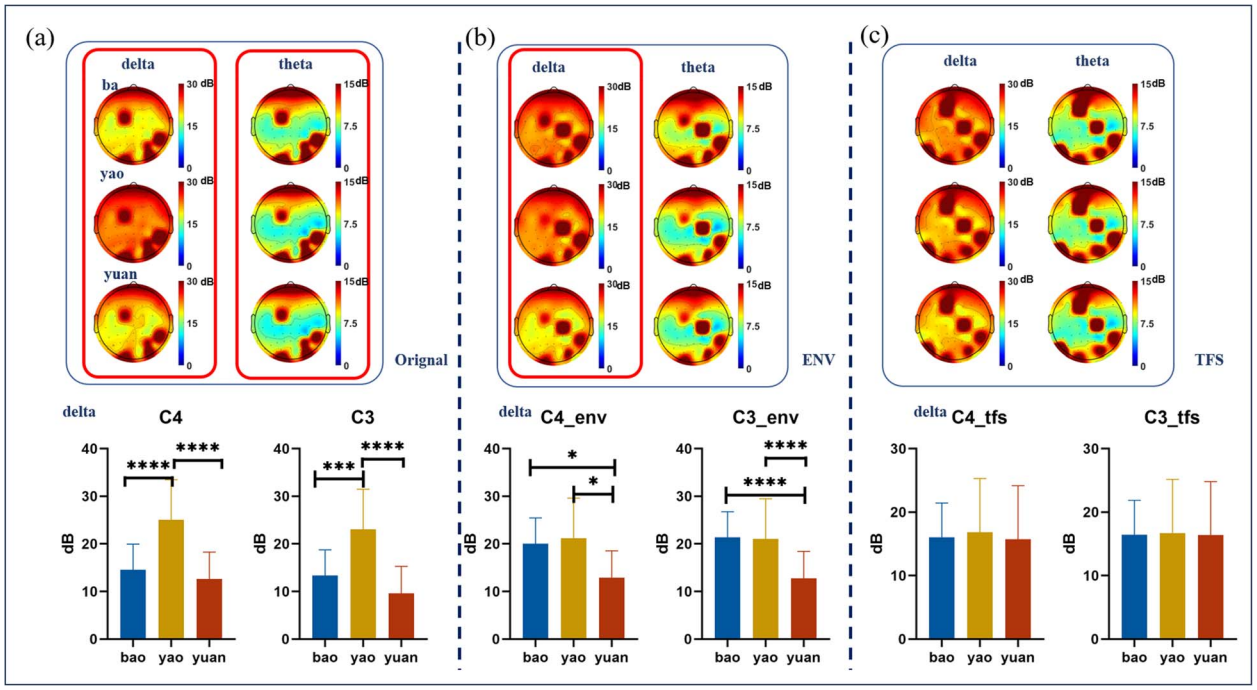

Fig. 7. The perceived PSD of the syllables at different levels of speech. (a) Shows syllable PSD results of the original speech. Different syllables had statistical differences in the delta and theta band, and the C3 and C4 had more significant statistical differences. (b) Shows the result of the syllable perceived PSD of the ENV. Different syllables had statistical differences in the delta band, and the C3 had more significant statistical differences, indicating that it tended to lateralization. (c) Shows the results of syllable perceived PSD of TFS, and there was no statistical difference between different syllables. (*$P < 0.05$, ***$P < 0.001$, ****$P < 0.0001$).

the statistical analysis of paired t-test. From Fig. 9(c), we could see no statistical difference between original speech and ENV in microstate 2–4 in syllable perception, but there were statistical differences in the main characteristic parameters of original speech and TFS, ENV, and TFS in microstate 3 and 4. Microstate 3 and 4 of TFS was significantly smaller than the original and ENV in terms of GEV, duration, and coverage, indicating that microstate 3 and 4 could be used as the main brain topographic map features of syllable perception experiment (Zhang et al. 2023). The specific statistical values of the microstate parameters of syllable perception were shown in Table 2. From the microstate conclusion and the behavioral results, we could infer that syllable perception mainly depended on the temporal envelope of speech.

The same analysis method was applied to the tone perception EEG data. In total, 9 best equivalent topographic maps were found, as shown in Fig. 10(a). We also look at the 100–400 ms microstate change under three conditions. We mainly analyzed microstates 3–6 and three main characteristic parameters and performed statistical analysis of paired t-tests. From Fig. 10(c), we could conclude that in tone perception, original speech, and TFS were only statistically significant in microstate 5 GEV and microstate 6 duration and coverage, and the other nine characteristic parameters were not statistically different. The specific statistical values of the microstate parameters of T3 perception were shown in Table 3. According to the behavioral and PSD results, the original speech and TFS were consistent in tone perception, therefore, the microstate brain topographic map 3–5 was mainly considered as the main characteristic parameter in tone perception. In tone perception, the three main parameters of ENV in microstates 3, 4, and 5 were significantly reduced compared with the original speech and TFS, which means that the speech information provided by

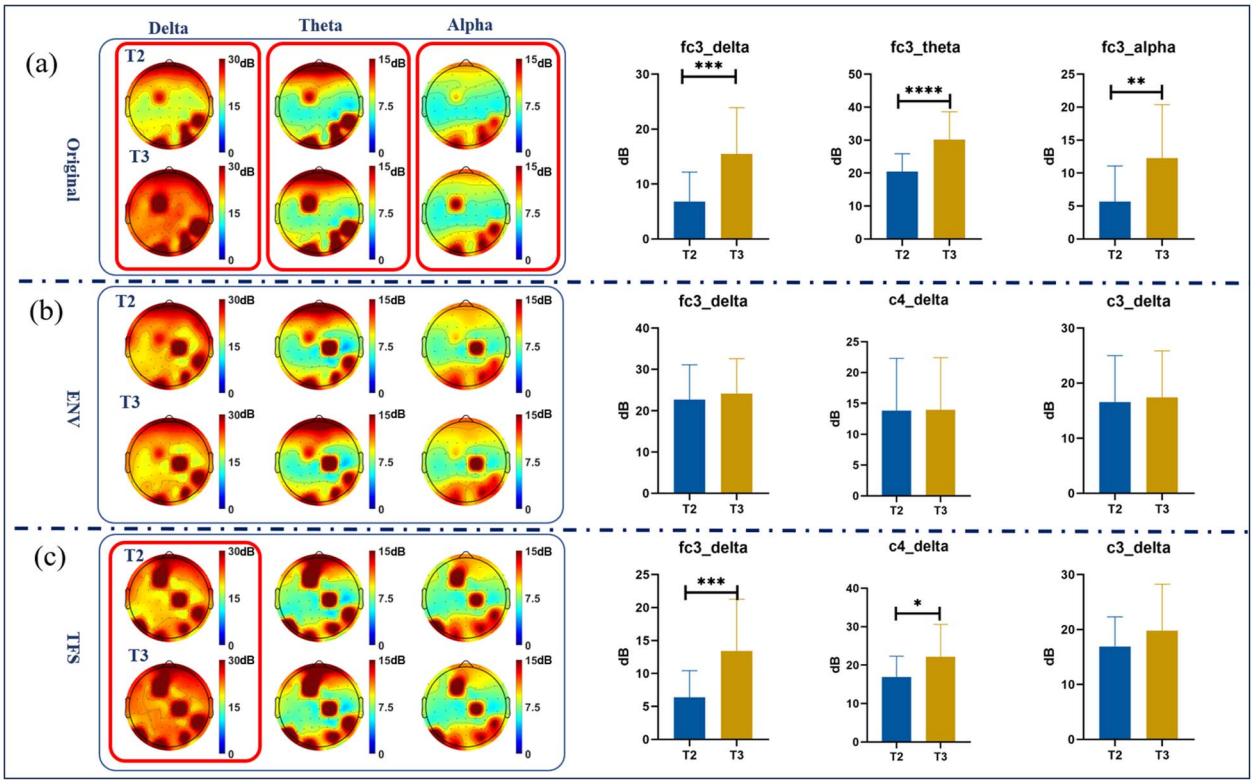

Fig. 8. Perceived PSD results from T2 and T3. (a) Shows the tone perception PSD results of the original speech. Different tones had statistical differences in delta, theta, and alpha band, and FC3 showed more significant statistical differences. (b) Shows the results of tone perception PSD of the ENV, and there was no statistical difference between different tones. (c) Shows the results of the PSD of tone perception in the TFS. The tone perception showed statistical differences in the delta frequency band and the C4 and FC3. (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$).

Table 2. Parametric statistical values of syllable perception microstates.

Syllable		Ori vs ENV	Ori vs TFS	ENV vs TFS
Gev(%)	MS2	0.5615	0.3472	0.1425
	MS3	0.8170	0.0010	0.0048
	MS4	0.5601	<0.0001	<0.0001
Duration(ms)	MS2	0.3050	0.2985	0.0396
	MS3	0.4335	0.0122	0.0047
	MS4	0.2332	<0.0001	<0.0001
Coverage(%)	MS2	0.5748	0.8476	0.5023
	MS3	0.7033	0.0107	0.0246
	MS4	0.2616	<0.0001	<0.0001

the speech envelope was not enough for the brain to make a tone judgment or the speech information provided may make the subject quickly judge it as T1, resulting in tone perception errors. Therefore, we could infer that tone perception mainly depends on the TFS of speech, and the speech envelope plays a certain auxiliary role based on the combination of behavioral, PSD, and microstate parameters.

Discussion

Auditory early cognitive process

The perception of Mandarin involves the processing of both tonal and syllabic information. However, it has always been debated whether these two types of information were fundamentally different or involved similar cognitive processes. Previous studies have investigated the difference between acoustic information and speech information from the perspective of the laterality of the brain (Xi et al. 2010; Zhang et al. 2011; Zhang et al. 2012a; Kayser et al. 2015). This study investigated the early auditory effects of tones and syllables in the processing of native Mandarin through the auditory chimeras. The results of this study showed that the brain response to Mandarin syllables and tones was different. In the ERP literature, peak latency is considered a temporal process of neural processing (Duncan et al. 2009). The influence of syllables and tones on the P2 peak latency of AEP was statistically significant, indicating that the neural processing process of different speech materials was different. Thus, the P2 peak latency could be a biomarker of Mandarin speech perception. Previous studies have shown that tonal and syllabic information could be processed in parallel at the attentional and pre-attentional stages. The results of this study suggested that the processing time of syllable information might be slightly earlier than that of tone information, which was consistent with previous research (Xi et al. 2010; Zhang et al. 2012a; Zhang et al. 2012b; Deroche et al. 2019).

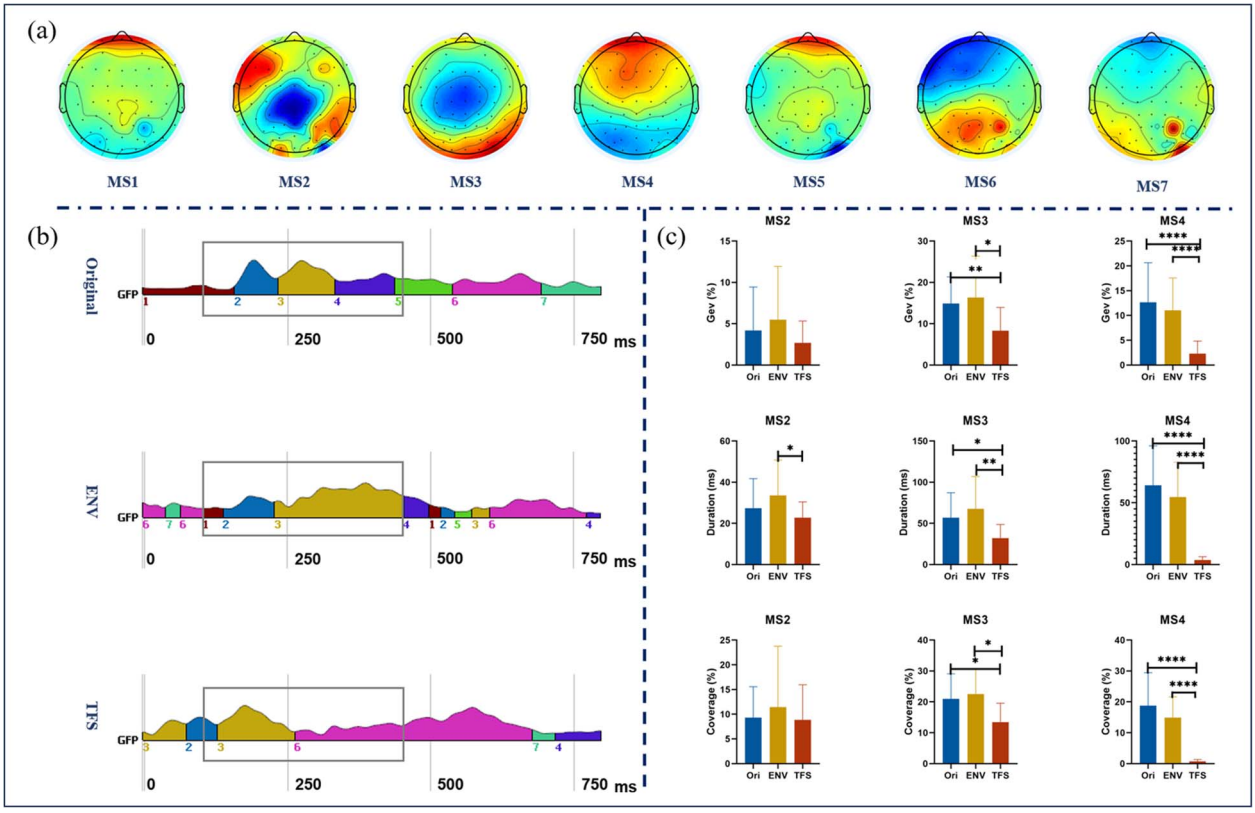

Fig. 9. Results of the microstate sequence of syllable perception at three acoustic conditions. (a) The best clustering template brain map of syllables at three acoustic conditions, showing the spatial characteristics of brain regions in each state. (b) The microstate time series characteristics of the same syllable (Yao) at three acoustic conditions. (c) The statistical results of three parameters (GEV, duration, and coverage) of microstate 2–4 of the syllable (Yao) task state. (*$P < 0.05$, **$P < 0.01$, ****$P < 0.0001$). The gray rectangular box was the microstate analysis area.

Table 3. Parametric statistical values of T3 perception microstates.

Syllable		Ori vs ENV	Ori vs TFS	ENV vs TFS
Gev(%)	MS3	0.0015	0.4410	0.0297
	MS4	0.0031	0.1611	0.0805
	MS5	<0.0001	0.0098	0.0498
	MS6	0.4885	0.1146	0.3472
Duration(ms)	MS3	0.0381	0.9986	0.0150
	MS4	0.1596	0.5280	0.3425
	MS5	0.0012	0.1584	0.0358
	MS6	0.9332	0.264	0.0014
Coverage(%)	MS3	0.0872	0.3552	0.2868
	MS4	0.0270	0.8565	0.0104
	MS5	0.0018	0.3871	0.0595
	MS6	0.3318	0.0301	0.0022

The role of delta and theta bands in Mandarin recognition

From the behavioral and EEG results, we could see that tone information was mainly carried by a TFS, while ENV mainly perceived syllable information. Recent studies have shown that cortical speech tracking in the theta frequency band was a significant predictor of speech intelligibility, while cortical speech tracking in the delta frequency band was the most relevant for speech understanding (Kosem and Van 2017; Etard and Reichenbach 2019). Our results also showed that the delta band of neural coherence was stronger during speech perception. We also found that the activation of brain areas for native Mandarin speakers was mainly temporal and prefrontal, as shown in Figs. 6 and 7 (C3, C4, and FC3 had statistical significance). The above results were consistent with the conclusions of previous studies (Ding et al. 2016; Broderick et al. 2018; Molinaro and Lizarazu 2018). They determined the role of the theta frequency band in faster neural activity in syllable tracking (Di Liberto et al. 2015). Continuous speech studies showed that the time scale of syllables was in the range of 4–8 Hz, and this frequency was consistent with theta rhythm. Indeed, theta cortical entrainment of TFS within this range was reported to be distinct from an envelope in a recent study (also using ENV and TFS decoupled stimuli) (Teng et al. 2019). In addition, the behavioral results show that TFS helps improve the recognition of speech whose envelope was temporally distorted. Although we found that the alpha frequency band is

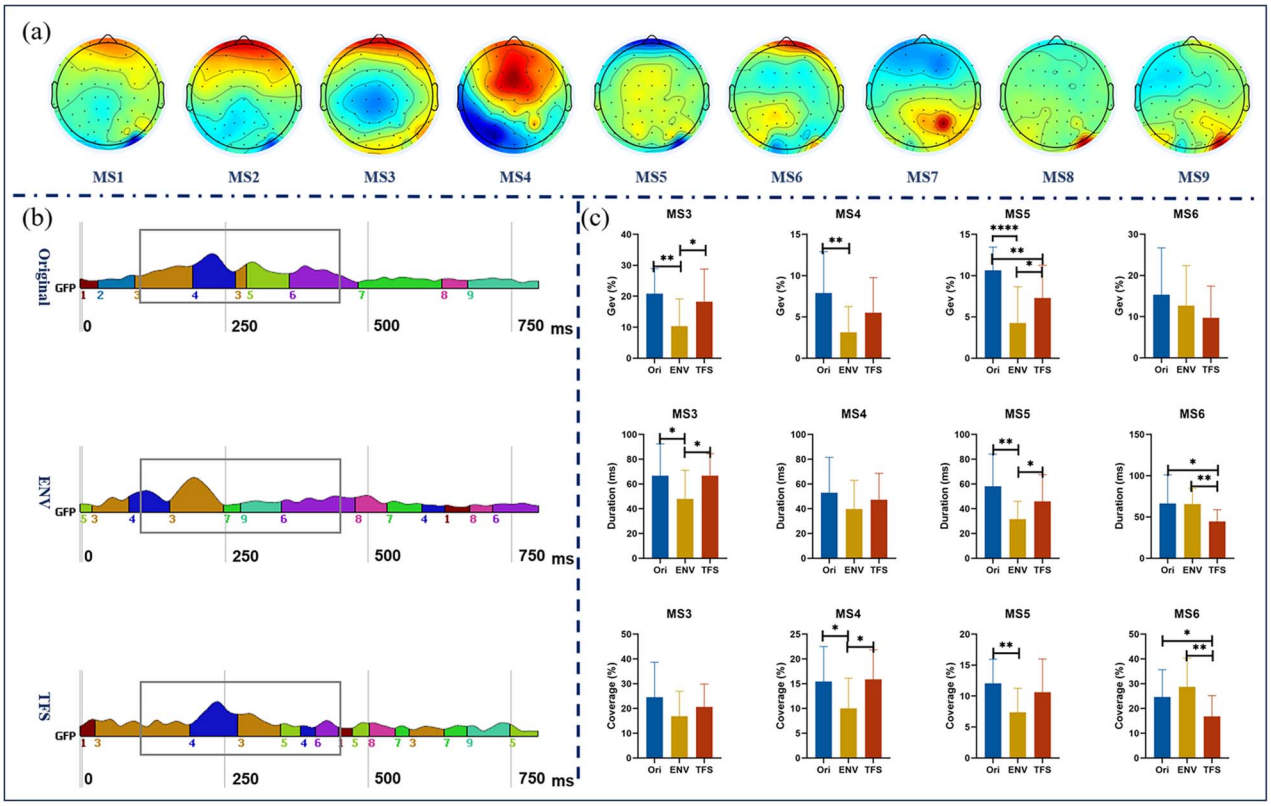

Fig. 10. Results of T3 perception microstate sequences at three acoustic conditions. (a) The best clustering template brain map of three tones at three acoustic conditions, representing the spatial characteristics of brain regions in each state. (b) The microstate time series characteristics of T3 at three acoustic conditions. (c) The statistical results of the three parameters (GEV, duration, and coverage) of the task-state microstate 2–4 of T3. (*$P < 0.05$, **$P < 0.01$, ****$P < 0.0001$). The gray rectangular box was the microstate analysis area.

statistically significant from the PSD of the original speech in the speech perception experiment paradigm, there was no significant difference in the speech perception experiment under the acoustic conditions of ENV and TFS. It might be that the power of the alpha frequency band changed but has no relationship with speech tracking (Khanna et al. 2015; Etard and Reichenbach 2019). When analyzing the results of tone perception in the three acoustic conditions, we found that the main contribution of tone perception was the delta frequency band, and the main brain area was the temporal lobe. Our statistical analysis showed that tones were biased to the right hemisphere and syllables to the left hemisphere during early auditory processing, which was consistent with the work of Luo et al. (2006).

The influence of TFS and ENV on Mandarin perception

Consistent with previous studies, subjects in our behavioral study could recognize tones with 60% accuracy using only ENV (without any spectral content) (Xu and Pfingst 2003). The behavioral study results showed that tone recognition accuracy was >90% when only the TFS information was used. The above results showed that TFS rather than ENV dominated tone perception. Previous studies have shown that the performance difference between ENV and TFS was more significant in noise. Our AEP, PSD, and microstate results explored from an EEG perspective that the main brain areas for tone perception were the prefrontal and temporal brain areas, and the main EEG rhythm was the delta band. The above results suggested that the neural processing time of tone may be ~200 ms after speech stimuli, i.e. the peak latency of the P2 component. From the results of the PSD and the microstates of the three acoustic conditions, we could see that the perception of speech was mainly dependent on the TFS, and the brain response time was mainly focused between 200 and 400 ms.

The superior temporal gyrus was an auditory association region whose activity was based on the spectral time representation of speech. It has shown a high degree of tuning to speech features (Chang et al. 2010; Mesgarani et al. 2014; Norman-Haignere and McDermott 2018). Some studies reported that cortical envelope tracking had shown sensitivity to speech intelligibility, i.e. syllable intelligibility (Peelle et al. 2013; Vanthornhout et al. 2018). However, other studies have not found this (Howard and Poeppel 2010). Researchers paid particular attention to the response of the cerebral cortex to higher-level linguistic features (including lexical and semantic levels). They often found a close relationship between attention and speech understanding (Howard and Poeppel 2010; Brodbeck et al. 2018; Broderick et al. 2018). Attention might play a role in our research, and subjects were required to focus on listening to the speech material and make important responses quickly. Our PSD and microstate results showed that syllable perception mainly depended on delta rhythm (strong speech envelope correlation) and prefrontal and parietal neural responses.

In summary, syllable perception depends mainly on ENV, and tone perception depends mainly on TFS. We could see from the microstates that the TFS in syllable perception lacks microstates 4 and 5 due to the subjects' inability to discriminate syllables well. The tone perception was not ideal in the same tone perception because the ENV microstate 4 was 70 ms ahead of time, and microstate 5 was missing. We provided reference and EEG support for the subsequent auditory brain-computer interface based on

Mandarin stimuli or cochlear encoder strategy by identifying the main EEG features of syllable and tone perception.

Acknowledgments

We thank to all the participants' for their cooperation in this study.

CRediT statements

Guangjian Ni (Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Validation, Writing—original draft), Zihao Xu (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft), Yanru Bai (Conceptualization, Methodology, Writing—original draft), Qi Zheng (Data curation, Validation, Visualization), Ran Zhao (Data curation, Visualization), Yubo Wu (Validation, Visualization), and Dong Ming (Conceptualization, Methodology, Project administration)

Funding

This work was supported by grants from the Key Technologies Research and Development Program of China (No. 2022YFF1202400) and the National Natural Science Foundation of China (81971698).

Conflict of interest statement: None declared.

Declaration of competing interest: The authors declare no competing interests.

Data and code availability

The data supporting this study's findings are available from the corresponding authors upon reasonable request. EEG data analyses were performed in MATLAB, the freely available toolbox EEGLAB, and Cartool64. The software code supporting this study's findings is available from the corresponding authors upon reasonable request.

References

Amaro E, Barker GJ. Study design in MRI: basic principles. *Brain Cogn*. 2006:60(3):220–232.

Apoux F, Healy EW. A glimpsing account of the role of temporal fine structure information in speech recognition. *Adv Exp Med Biol*. 2013:787:119–126.

Britz J, Michel CM. State-dependent visual processing. *Front Psychol*. 2011:2:00370.

Britz J, Díaz Hemandez L, Ro T. EEG-microstate dependent emergence of perceptual awareness. *Front Behav Neurosci*. 2014:8:00163.

Brodbeck C, Hong LE, Simon JZ. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr Biol*. 2018:28:3976–3983.

Broderick MP, Anderson AJ, di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr Biol*. 2018:28(5):803–809.e3.

Brohl F, Kayser C. Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *NeuroImage*. 2021:233:117958.

Chang EF, Rieger JW, Johnson K, et al. Categorical speech representation in human superior temporal gyrus. *Nat Neurosci*. 2010:13:1428–U1169.

Chao YR. *A grammar of spoken Chinese*. Oakland, CA: The University of California Press; 1965.

Deroche MLD, Lu HP, Lin YS, et al. Processing of acoustic information in lexical tone production and perception by pediatric cochlear implant recipients. *Front Neurosci*. 2019:13:00639.

Di Liberto GM, O'Sullivan JA, Lalor EC. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol*. 2015:25:2457–2465.

Ding N, Simon JZ. Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci*. 2014:8:00311.

Ding N, Melloni L, Zhang H, et al. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci*. 2016:19:158–164.

Duncan CC, Barry RJ, Connolly JF, et al. Event-related potentials in clinical research: guidelines for eliciting, recording and quantifying mismatch negativity, P300 and N400. *Clin Neurophysiol*. 2009:120:1883–1908.

Etard O, Reichenbach T. Neural speech tracking in the theta and in the Delta frequency band differentially encode clarity and comprehension of speech in noise. *J Neurosci*. 2019:39:5750–5759.

Gandour J, Wong D, Lowe M, et al. A cross-linguistic fMRI study of spectral and temporal cues underlying phonological processing. *J Cogn Neurosci*. 2002:14:1076–1087.

Goswami U. Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Ann N Y Acad Sci*. 2019:1453(1):14137.

Gui P, Jiang Y, Zang D, Qi Z, Tan J, Tanigawa H, Jiang J, Wen Y, Xu L, Zhao J, et al. Assessing the depth of language processing in patients with disorders of consciousness. *Nat Neurosci*. 2020:23(6):761–770.

Ho A, Boshra R, Schmidtke D, Oralova G, Moro AL, Service E, Connolly JF. Electrophysiological evidence for the integral nature of tone in mandarin spoken word recognition. *Neuropsychologia*. 2019:131:325–332.

Hopkins K, Moore BC. The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *J Acoust Soc Am*. 2009:125:442–446.

Hopkins K, Moore BC, Stone MA. Effects of moderate cochlear hearing loss on the ability to ben-efit from temporal fine structure information in speech. *J Acoust Soc Am*. 2008:123(2):1140–1153.

Howard MF, Poeppel D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol*. 2010:104:2500–2511.

Joris PX. Responses to amplitude-modulated tones in the auditory nerve of the cat. *J Acoust Soc Am*. 1992:91:215–232.

Kayser SJ, Ince RA, Gross J, Kayser C. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J Neurosci*. 2015:35(44):14691–14701.

Khanna A, Pascual-Leone A, Michel CM, Farzan F. Microstates in resting-state EEG: current status and future directions. *Neurosci Biobehav Rev*. 2015:49:105–113.

Kim SG, Richter W, Ugurbil K. Limitations of temporal resolution in functional MRI. *Magn Reson Med*. 1997:37:631–636.

Kindler J, Hubl D, Strik WK, Dierks T, Koenig T. Resting-state EEG in schizophrenia: auditory verbal hallucinations are related to shortening of specific microstates. *Clin Neurophysiol*. 2011:122(6):1179–1182.

Klein D, Zatorre RJ, Milner B, Zhao V. A cross-linguistic PET study of tone perception in mandarin Chinese and English speakers. *NeuroImage*. 2001:13(4):646–653.

Kong YY, Zeng FG. Temporal and spectral cues in mandarin tone recognition. *J Acoust Soc Am*. 2006:120:2830–2840.

Kosem A, Van WV. Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. *Lang Cogn Neurosci*. 2017:32:536–544.

Lehmann D, Faber PL, Galderisi S, Herrmann WM, Kinoshita T, Koukkou M, Mucci A, Pascual-Marqui RD, Saito N, Wackermann J, et al. EEG microstate duration and syntax in acute, medication-naive, first-episode schizophrenia: a multi-center study. *Psychiatry Res*. 2005:138(2):141–156.

Li Y, Tang C, Lu J, Wu J, Chang EF. Human cortical encoding of pitch in tonal and non-tonal languages. *Nat Commun*. 2021:12(1):1161.

Liu JY, Xu J, Zou GY, He Y, Zou Q, Gao JH 2020. Reliability and individual specificity of EEG microstate characteristics. *Brain Topogr* 33:101007, 4, 438, 449.

Luo H, Ni JT, Li ZH, Li XO, Zhang DR, Zeng FG, Chen L. Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proc Natl Acad Sci USA*. 2006:103(51):19558–19563.

Meng QL, Zheng NH, Li X. Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants. *J Acoust Soc Am*. 2016:139:301–310.

Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic feature encoding in human superior temporal gyrus. *Science*. 2014:343(6174):1006–1010.

Michel CM, Koenig T. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *NeuroImage*. 2018:180:577–593.

Mishra A, Englitz B, Cohen MX. EEG microstates as a continuous phenomenon. *NeuroImage*. 2020:208:116454.

Molinaro N, Lizarazu M. Delta (but not theta)-band cortical entrainment involves speech-specific processing. *Eur J Neurosci*. 2018:48:2642–2650.

Norman-Haignere SV, McDermott JH. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol*. 2018:16:e2005127.

Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex*. 2013:23:1378–1387.

Prinsloo KD, Lalor EC. General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras. *J Neurosci*. 2022:42:7782–7798.

Rose JE, Brugge JF, Anderson DJ, et al. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J Neurophysiol*. 1967:30:769–793.

Roy AT, Carver C, Jiradejvong P, et al. Musical sound quality in cochlear implant users: a comparison in bass frequency perception between fine structure processing and high-definition continuous interleaved sampling strategies. *Ear Hear*. 2015:36:582–590.

Smith Z, Delgutte B, Oxenham JA. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*. 2002:416:87–90.

Teng X, Cogan GB, Poeppel D. Speech fine structure contains critical temporal cues to support speech segmentation. *NeuroImage*. 2019:202:116152.

Vandali A, Dawson P, Au A, Yu Y, Brown M, Goorevich M, Cowan R. Evaluation of the optimized pitch and language strategy in cochlear implant recipients. *Ear Hear*. 2019:40(3):555–567.

Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T. Speech intelligibility predicted from neural entrainment of the speech envelope. *J Assoc Res Otolaryngol*. 2018:19(2):181–191.

Wang S, Liu B, Zhang H, Dong R, Mannell R, Newall P, Chen X, Qi B, Zhang L, Han D. Mandarin lexical tone recognition in sensorineural hearing-impaired listeners and cochlear implant user. *Acta Otolaryngol*. 2013:133(1):47–54.

Wei Y, Liang X, Guo X, Wang X, Qi Y, Ali R, Wu M, Qian R, Wang M, Qiu B, et al. Brain hemispheres with right temporal lobe damage swap dominance in early auditory processing of lexical tones. *Front Neurosci*. 2022:16:909796.

Wilson BW, Finley CC, Lawson D. Better speech recognition with cochlear implants. *Nature*. 1991:12:236–238.

Wong PCM, Parsons LM, Martinez M, Diehl RL. The role of the insular cortex in pitch pattern perception: the effect of linguistic contexts. *J Neurosci*. 2004:24(41):9153–9160.

Xi J, Zhang L, Shu H, Zhang Y, Li P. Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience*. 2010:170(1):223–231.

Xu L, Pfingst BE. Relative importance of temporal envelope and fine structure in lexical-tone perception. *J Acoust Soc Am*. 2003:114:3024–3027.

Yu K, Wang R, Li L, Li P. Processing of acoustic and phonological information of lexical tones in mandarin Chinese revealed by mismatch negativity. *Front Hum Neurosci*. 2014:8:729.

Zhang L, Xi J, Xu G, Shu H, Wang X, Li P. Cortical dynamics of acoustic and phonological processing in speech perception. *PLoS One*. 2011:6(6):e20963.

Zhang L, Xi J, Wu H, Shu H, Li P. Electrophysiological evidence of categorical perception of Chinese lexical tones in attentive condition. *Neuroreport*. 2012a:23(1):35–39.

Zhang Y, Zhang L, Shu H, Xi J, Wu H, Zhang Y, Li P. Universality of categorical perception deficit in developmental dyslexia: an investigation of mandarin Chinese tones. *J Child Psychol Psychiatry*. 2012b:53(8):874–882.

Zhang C, Yang Y, Han S, Xu L, Chen X, Geng X, Bie L, He J. The temporal dynamics of large-scale brain network changes in disorders of consciousness: a microstate-based study. *CNS Neurosci Ther*. 2023:29(1):296–305.