Behavioral/Cognitive

Attention Is Required for Knowledge-Based Sequential Grouping: Insights from the Integration of Syllables into Words

Nai Ding,[1,2,3,4,7] **Xunyi Pan,**[5] **Cheng Luo,**[1] **Naifei Su,**[1] **Wen Zhang,**[1] **and Jianfeng Zhang**[1,6]

[1]College of Biomedical Engineering and Instrument Sciences, [2]Key Laboratory for Biomedical Engineering of Ministry of Education, [3]State Key Laboratory of Industrial Control Technology, [4]Interdisciplinary Center for Social Sciences, [5]School of International Studies, [6]Mental Health Center, School of Medicine, Zhejiang University, Hangzhou, China 310027, and [7]Neuro and Behavior EconLab, Zhejiang University of Finance and Economics, Hangzhou, China 310027

How the brain groups sequential sensory events into chunks is a fundamental question in cognitive neuroscience. This study investigates whether top–down attention or specific tasks are required for the brain to apply lexical knowledge to group syllables into words. Neural responses tracking the syllabic and word rhythms of a rhythmic speech sequence were concurrently monitored using electroencephalography (EEG). The participants performed different tasks, attending to either the rhythmic speech sequence or a distractor, which was another speech stream or a nonlinguistic auditory/visual stimulus. Attention to speech, but not a lexical-meaning-related task, was required for reliable neural tracking of words, even when the distractor was a nonlinguistic stimulus presented cross-modally. Neural tracking of syllables, however, was reliably observed in all tested conditions. These results strongly suggest that neural encoding of individual auditory events (i.e., syllables) is automatic, while knowledge-based construction of temporal chunks (i.e., words) crucially relies on top–down attention.

Key words: attention; entrainment; speech; syllables; words

Significance Statement

Why we cannot understand speech when not paying attention is an old question in psychology and cognitive neuroscience. Speech processing is a complex process that involves multiple stages, e.g., hearing and analyzing the speech sound, recognizing words, and combining words into phrases and sentences. The current study investigates which speech-processing stage is blocked when we do not listen carefully. We show that the brain can reliably encode syllables, basic units of speech sounds, even when we do not pay attention. Nevertheless, when distracted, the brain cannot group syllables into multisyllabic words, which are basic units for speech meaning. Therefore, the process of converting speech sound into meaning crucially relies on attention.

Introduction

Sequentially grouping events into temporal chunks is a fundamental function of the brain (Lashley, 1951; Gavornik and Bear, 2014) and is especially important for audition. During speech comprehension, for example, sequential grouping occurs hierarchically, with syllables being grouped into words and words being grouped into phrases, sentences, and discourses. Similarly, during music perception, musical notes are hierarchically grouped into meters and phrases (Patel, 2008). Whether auditory sequential grouping requires attention is under debate (Snyder et al., 2006; Shinn-Cunningham, 2008; Shinn-Cunningham et al., 2017). On the one hand, it has been hypothesized that top–down attention is required for sequential grouping, especially for a complex auditory scene consisting of multiple auditory sequences. Evidence indicates that attention strongly modulates neural and behavioral responses to sound sequences (Carlyon et al., 2001; Fritz et al., 2007; Shamma et al., 2011; Lu et al., 2017). Research on visual object recognition has also suggested that top–down attention is required for the binding of simultaneously presented features, e.g., color and shape (Treisman and Gelade, 1980). On the other hand, many neurophysiological studies have shown that the brain is sensitive to temporal regularities in sound even when the sound is not attended (Näätänen et al., 2007; Sussman et al., 2007; Barascud et al., 2016), suggesting that primitive analyses of temporal sequences may occur as an automatic process (Fodor, 1983).

Received Sept. 11, 2017; revised Nov. 8, 2017; accepted Dec. 5, 2017.

Author contributions: N.D. designed research; X.P., C.L., N.S., W.Z., and J.Z. performed research; N.D. contributed unpublished reagents/analytic tools; N.D., C.L., N.S., and W.Z. analyzed data; N.D., X.P., and J.Z. wrote the paper.

This work was supported by the National Natural Science Foundation of China (Grants 31771248 and 31500873 to N.D.), by the Zhejiang Provincial Natural Science Foundation of China (Grant LR16C090002 to N.D.), by Fundamental Research Funds for the Central Universities (N.D.), and by research funding from the State Key Laboratory of Industrial Control Technology, Zhejiang University (N.D.). We thank Yanchao Bi, Inyong Choi, Huan Luo, Lucia Melloni, Jonas Obleser, David Poeppel, Jonathan Simon, and Elana Zion-Golumbic for helpful comments on earlier versions of this manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Nai Ding, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China 310027. E-mail: ding_nai@zju.edu.cn.

DOI:10.1523/JNEUROSCI.2606-17.2017

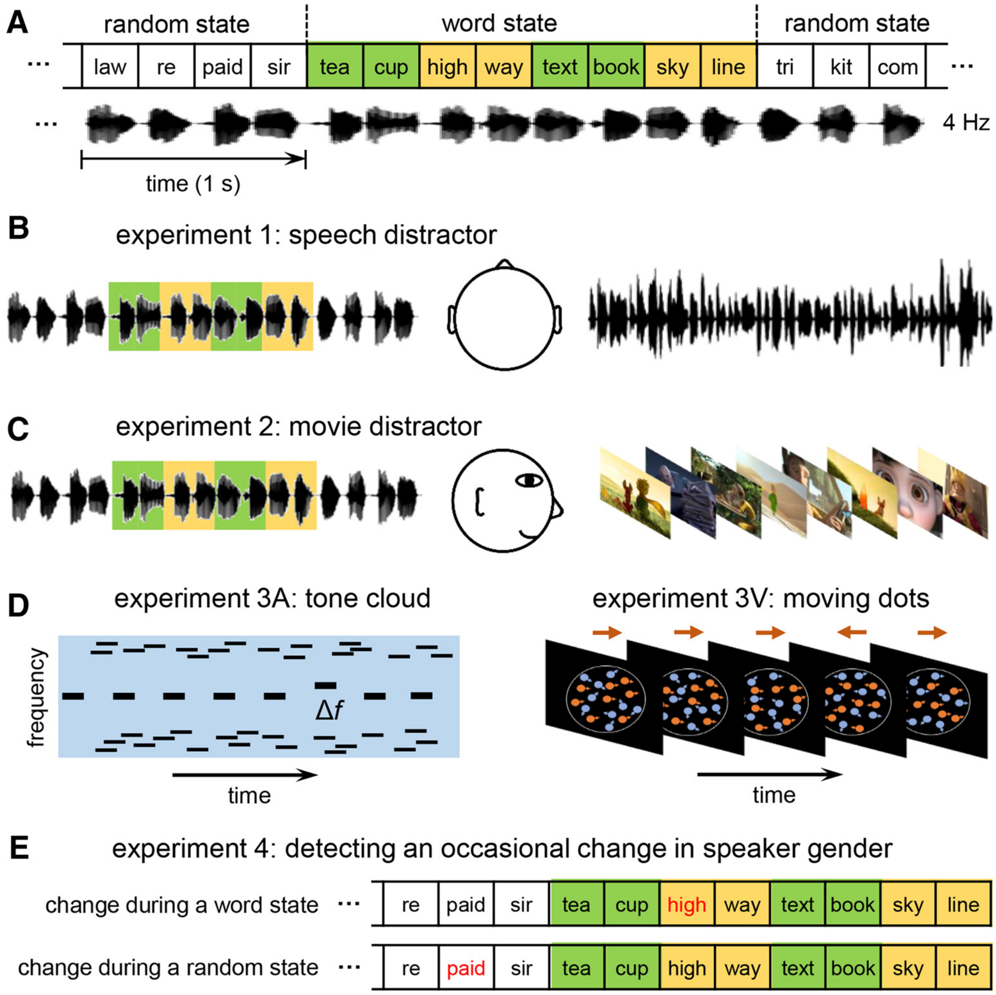

A
random state
word state
random state
...
law
re
paid
sir
tea
cup
high
way
text
book
sky
line
tri
kit
com
...
4 Hz
time (1 s)
B
experiment 1: speech distractor
C
experiment 2: movie distractor
D
experiment 3A: tone cloud
experiment 3V: moving dots
frequency
Δf
time
time
E
experiment 4: detecting an occasional change in speaker gender
change during a word state
...
re
paid
sir
tea
cup
high
way
text
book
sky
line
change during a random state
...
re
paid
sir
tea
cup
high
way
text
book
sky
line

Figure 1. Experiment design. ***A***, Structure of the isochronous syllable sequence, which alternated between word states and random states. The syllables were presented at a constant rate of 4 Hz and therefore the bisyllabic words were presented at 2 Hz. English syllables are shown in the figure for illustrative purposes. Chinese syllables and words were used in the experiments. ***B***, In Experiment 1, the isochronous syllable sequence and a competing spoken passage were simultaneously presented to different ears. The participants attended to different ears in different experimental blocks. ***C***, In Experiment 2, the listeners either attended to the isochronous syllable sequence (presented to both ears) or watched a movie while passively listening to the syllable sequence. ***D***, The auditory and visual distractor used in Experiment 3. The auditory distractor consisted of a 3 Hz tone sequence embedded in a tone cloud. The auditory distractor and the isochronous syllable sequence were presented dichotically. The participants had to detect occasional frequency deviants in the tone sequence. The visual distractor consisted of orange and cyan dots moving in different directions. Dots of one color moved randomly while dots of the other color showed partly coherent motion. The participants had to detect occasional reversals in the direction of the coherent motion. ***E***, Experiment 4 presented the isochronous syllable sequence without any distractor. The participants had to detect occasional changes in the speaker gender, which could occur either during the word state or during the random state.

Sequential grouping is not a single computational module, which further complicates the discussion about how sequential grouping is modulated by attention. Sequential-grouping mechanisms include bottom–up primitive grouping and top–down schema-based grouping (Bregman, 1990). Bottom–up grouping depends on the similarity between sensory features (Micheyl et al., 2005; McDermott et al., 2011; Woods and McDermott, 2015), while top–down schema-based grouping relies on prior knowledge (Hannemann et al., 2007; Jones and Freyman, 2012; Billig et al., 2013). Both grouping mechanisms play important roles in auditory perception. For example, in spoken-word recognition, integrating acoustic features into phonemes and syllables can rely on acoustic continuity cues within a syllable (Shinn-Cunningham et al., 2017), while integrating syllables into words crucially relies on lexical knowledge, i.e., the knowledge about which syllable combinations constitute valid words (Mattys et al., 2009; Cutler, 2012). Most previous studies focus on how attention modulates primitive sequential grouping while relatively little is known about how attention modulates schema-based grouping. Using four experiments, the current study fills this gap by studying the neural processes underlying the knowledge-based grouping of syllables into words (Fig. 1).

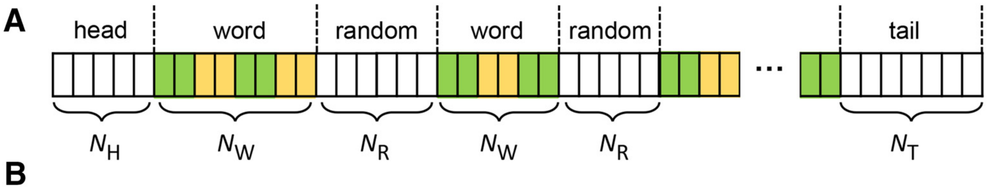
Materials and Methods

Participants

Fifty-two participants took part in the study (18–27 years old; mean age, 23 years; 48% female). Each experiment included 14 participants. No experiment had >1 participant who took part in another experiment. All participants were graduate or undergraduate students at Zhejiang University, with no self-reported hearing loss or neurological disorders. The experimental procedures were approved by the Institutional Review Board of the Zhejiang University Interdisciplinary Center for Social Sciences. The participants provided written consent and were paid for taking part.

Word materials

The study used 160 animate words and 160 inanimate words, all of which are bisyllabic Chinese words. Animate words included animals ($N = 40$;



B

	M	N_H	N_W	N_R	N_T
experiment 1	3	8+U(0,3)	12+2U(0,1)	7+U(0,1)	$72\text{-}N_H\text{-}\Sigma_m(N_W+N_R)$
experiments 2-4	5	5+G(0.6)	6+2G(0.3)	7+G(0.6)	7+G(0.6)

U(a,b): discrete uniform distribution between a and b
G(p): geometric distribution: $P(X = k) = (1-p)p^k$
E(U(0,3)) = 1.5, E(U(0,1)) = 0.5, E(G(0.3))=1.5, and E(G(0.6))=0.43

Figure 2. Structure of the isochronous syllable sequence in each experiment. ***A***, The sequence alternated between random states and word states M times in each trial. At the beginning and end of each trial, N_H and N_T random syllables were presented. ***B***, Statistical distribution of the number of syllables in each state.

e.g., monkey, dolphin), plants ($N = 40$; e.g., lemon, carrot), occupations ($N = 48$; e.g., doctor, doorman), and names of well-known people in history ($N = 32$; e.g., Bai Li, a famous poet in Tang dynasty). Inanimate words include nonliving things ($N = 80$; e.g., teacup, pencil) and places ($N = 80$; e.g., Beijing, Zhejiang).

Stimuli

Isochronous syllable sequence. The main stimulus is an isochronous syllable sequence (Fig. 1*A*). All syllables were independently synthesized using the Neospeech synthesizer (http://www.neospeech.com/; the male voice, Liang) and were adjusted to the same intensity and the same duration, i.e., 250 ms (Ding et al., 2016). The syllable sequence alternated between a word state and a random state (Fig. 2*A*). The number of syllables in each state and the number of word states in each stimulus (i.e., M) are shown in Figure 2*B*. Each sequence started and ended with a random state to reduce the possibility that words might be noticed at the beginning and end of each stimulus when the syllable sequence was not attended to. No word appeared twice in a trial and there was no immediate repetition of any syllable. Words in the same word state were either all animate words or all inanimate words, and the animacy of each word state was randomly chosen. The participants were never told how many word states might appear in a trial.

In Experiment 1, the number of syllables in the word/random states was randomized using a uniform distribution so that the alternation between states was not regular while the total duration could be easily controlled. Experiments 2–4 used the same set of isochronous syllable sequences. To further reduce the predictability of the onset/offset of each word state, the number of syllables in the word and random states was subject to a geometric distribution so that the participants could not predict when state transitions would occur.

Spoken message distractors in Experiment 1. In Experiment 1, an isochronous syllable sequence and a competing spoken passage were dichotically presented (Fig. 1*B*). The ear to which each stimulus was presented was counterbalanced across participants. The competing spoken passages (chosen from the *Syllabus for Mandarin Proficiency Tests*) were time compressed by a factor of 2.5 and gaps longer than 30 ms were shortened to 30 ms. Long acoustic pauses were removed since the listeners might shift their attentional focus during the long pauses. In each trial, 19 s of spoken passages were presented and the duration of each syllable sequence was set to 18 s, i.e., 72 syllables. The competing spoken passage started 1 s before the syllable sequence so that the syllable sequence was less likely to be noticed when the listeners focused on the spoken passage.

Auditory and visual distractors in Experiment 3. Experiment 3 was divided into an auditory-distractor condition and a visual-distractor condition (Fig. 1*C*). The auditory distractor consisted of a 3 Hz tone sequence embedded in a tone cloud. Each tone was 75 ms in duration and its onset and offset were smoothed by a 10 ms cosine ramp. The 3 Hz tone sequence had a fixed frequency, f_T, which was uniformly distributed between 512 and 1024 Hz in a log frequency scale. The tone cloud consisted of 50 tones per second. The tone cloud and the 3 Hz tone sequence did not overlap in frequency. In the tone cloud, the tone frequency, f_C, was ≥0.5 octave higher or lower than f_T. Specifically, f_C was uniformly distributed in two two-octave-wide spectral regions: $[\log_2(f_T) - 2.5 < \log_2(f_C) < \log_2(f_T) - 0.5]$ and $[\log_2(f_T) + 0.5 < \log_2(f_C) < \log_2(f_T) + 2.5]$. In the stimulus, the tone cloud started 0.5 s after the onset of the 3 Hz tone sequence, allowing the participants to hear the first two tones in the 3 Hz sequence without any acoustic interference.

The visual distractor in Experiment 3 consisted of orange (RGB: 250, 200, 0) and cyan (RGB: 0, 200, 250) dots moving in a black background. On average, 136 dots appeared in a circular region (~10° visual angle in diameter), half of which were orange. The velocity of each dot was the vector sum of three components: a color-dependent component v_C, an item-specific component v_i, and an item-specific time-varying noise term $\varepsilon_i(t)$. A nonzero v_C leads to coherent motion across dots of the same color. The item-specific and time-varying components, however, lead to incoherent motion direction across dots.

The color-dependent component v_C was identical for dots of the same color but varied across trials. In each trial, v_C was zero for dots of one color and was $v_C = [A\cos\alpha, A\sin\alpha]$ for dots of the other color, where α denotes the motion direction and was uniformly distributed between 0 and 360°. The item-specific component v_i was independently generated for each dot. It was denoted as $v_i = [A\cos\varphi_i, A\sin\varphi_i]$ for dot i, where φ_i was uniformly distributed between 0 and 360°. The time-varying component $\varepsilon_i(t) = [A\cos\theta_i(t), A\sin\theta_i(t)]$ was independently generated for each dot i at each time point t and $\theta_i(t)$ was uniformly distributed between 0 and 360°. In the experiment, the participants sat ~60 cm away from the screen. The moving speed, A, is ~7° per second. The position and velocity of each dot was updated at the screen's refresh rate, i.e., 85 Hz.

Procedures and tasks

The study consisted of four experiments. Experiments 1–3 contained two blocks, which differed depending on the attentional focus of the participants.

Experiment 1. In the first block, listeners attended to the time-compressed spoken passage and answered comprehension questions after each trial. The comprehension questions were presented 1 s after the spoken passage and the participants had to give a verbal answer. After recording the answer, the person conducting the experiment pressed a key to continue the experiment. The next trial was played after an interval randomized between 1 and 2 s (uniform distribution) after the key press.

In the second block, participants had to focus on the syllable sequences and judge whether an additional word presented 1 s after the sequence offset appeared in the sequence. They pressed different keys to indicate whether the word appeared in the sequence or not. The next trial started after an interval randomized between 1 and 2 s (uniform distribution) after the key press. The same set of 50 trials (50 distinct spoken passages paired with 50 distinct syllable sequences) were presented in each block in a random order. The participants had their eyes closed when listening to the stimuli and had a break every 25 trials. The block in which participants attended to spoken passages was always run first, in case the participants might spontaneously shift their attention to the isochronous syllables after knowing there were words embedded in the sequence.

Experiment 2. A word-listening block and a movie-watching block were presented, the order of which was counterbalanced across participants. In the word-listening block, after each trial, participants had to press different keys to indicate whether they heard more animate words or more inanimate words. The participants were told that all words within the same word state belonged to the same category (i.e., animate or inanimate) and therefore they only had to indicate whether a trial had more animate word states or inanimate word states. Sixty trials were presented and the participants had a break after every 15 trials. Before the word-listening condition, the participants went through a practice section, in which they listened to two example sequences and did the same task. They received feedback during the practice session but not during the main experiment. The neural responses showed the same pattern whichever block was presented first and therefore the responses were averaged over all participants regardless of the presentation order.

In the movie-watching block, the participants watched a silent movie (*The Little Prince*) with Chinese subtitles. The syllable sequences were presented ~3 min after the movie started to make sure participants had already engaged in the movie-watching task. Sixty syllable sequences were presented in a randomized order, with the interstimulus-interval randomized between 1 and 2 s. The movie was stopped after all the 60 trials were presented. The participants had their eyes open in both blocks although no visual stimulus was presented in the word-listening block.

Experiment 3. An auditory-distractor block and a visual-distractor block were presented, the order of which was counterbalanced across participants. Before each block, the participants were told that they would hear a task-irrelevant speech signal that they should ignore. The isochronous syllable sequences used in both blocks were identical to those used in Experiment 2. In the auditory-distractor block, the auditory distractor and the isochronous syllable sequence were dichotically presented. The auditory distractor started 2 s before the onset of the syllable sequence and stopped the same time when the syllable sequence stopped. The participants had to detect occasional frequency deviants in the 3 Hz tone sequence embedded in a tone cloud (Fig. 1*D*). A spectral deviant was created by shifting the frequency of one tone up or down by two semitones. No deviants were presented in the first or the last 3 s of the stimulus. When two deviants occurred, they were separated by ≥1 s. In the 60 trials being presented, 30 trials contained a single frequency deviant and 15 trials contained two deviants.

In the visual block, the visual distractor started 1 s before the diotically presented isochronous syllable sequence and stopped 1 s after the offset of the syllable sequence. The participants had to detect occasional reversals in the direction of the coherent motion (Fig. 1*E*). A reversal in motion direction lasted for 350 ms. When the motion direction reversed twice, the two reversals were separated by ≥2 s. In both the auditory and visual blocks, the participants pressed different keys at the end of each trial based on whether they heard deviants or not, and the next trial started 1–2 s (uniform distribution) after the key press.

Experiment 4. Experiment 4 used the same set of stimuli used in Experiments 2 and 3, and the syllables were diotically presented. Ten additional trials were created, in which the voice of a randomly chosen syllable was changed using the change-gender function in Praat (Boersma and Weenink, 2017). The spectrum of the syllable was shifted up by a factor of 1.1 so that the perceived gender of the speaker changed from male to female. The position of the syllable with a female voice was uniformly distributed from 2 s after the stimulus onset to 2 s before the stimulus offset. The participants had to detect such a change in the speaker's gender. The gender-detection task did not require lexical processing and the purpose of the task was to test whether lexical processing might automatically occur without explicit task demand.

EEG recording and analysis

EEG responses were recorded using a 64-channel Biosemi ActiveTwo system. Additionally, four electrodes were used to record horizontal and vertical EOGs and two reference electrodes were placed at the left and right mastoids. The EEG recordings were low-pass filtered below 400 Hz and sampled at 2048 Hz. The EEG recordings were referenced to the average mastoid recording off-line and the horizontal and vertical EOG signals were regressed out. Since the study focused on word-rate and syllable-rate neural responses (2 and 4 Hz respectively), the EEG recordings were high-pass filtered above 0.7 Hz. Each analysis epoch was 9 s in duration, beginning 2 s before the onset of each word state. All epochs were averaged before the frequency-domain analysis.

In the frequency domain analysis, a discrete Fourier transform was applied to each EEG channel and each participant. The analysis window was 2 s in duration, resulting in a frequency resolution of 0.5 Hz. In Experiments 2–4, a single analysis window was used, which started from the word-state onset. In Experiment 1, since the word state is longer, two successive analysis windows were applied, with the first one starting from the word-state onset and the second starting from the offset of the first analysis window. The EEG spectrum was averaged over EEG channels and participants (and analysis windows in Experiment 1) by calculating the root-mean-square value.

In the time domain analysis of the word-rate response, the EEG responses were filtered ~2 Hz using a Hamming-window-based, linear-phase finite impulse response (FIR) filter. The filter impulse response duration was 1 s and the gain at 4 Hz was −45 dB. The linear delay caused by the FIR filter was compensated by shifting the filtered signal back in time. The instantaneous amplitude of the word-rate response was extracted using the Hilbert transform. The magnitude of the Hilbert transform was further low-pass filtered (−6 dB decay at 1 Hz, −35 dB decay at 2 Hz), and converted into a decibel amplitude scale.

Statistical test

This study used bias-corrected and accelerated bootstrap for all significance tests (Efron and Tibshirani, 1993). In the bootstrap procedure, all the participants were resampled with replacement 10^4 times. For the significance tests for the 2 and 4 Hz peaks in the response spectrum (Figs. 3–5), the response amplitude at the peak frequency was compared with the mean amplitude of the neighboring two frequency bins (corresponding to a 1 Hz width). In the interparticipant phase coherence test in Figure 5, 10^4 phase coherence values were generated based on the null distribution, i.e., uniform distribution of response phase across participants. If the experimentally measured phase coherence was $<N$ of the 10^4 randomly generated phase coherence values, its significance level was $(N + 1)/10^4$.

For the significance test for time intervals showing response amplitude differences (Fig. 6), the EEG waveform was averaged in the bootstrap procedure over all sampled participants and the instantaneous amplitude was then extracted. For paired comparisons, the two conditions being compared went through the same resampling procedure. For unpaired conditions, the significance level is v if the sample mean in one condition exceeds the 100-v/2 percentile (or falls below the v/2 percentile) of the distribution of the sample mean in the other condition.

Results

We used spoken-word processing as a paradigm to test how attention may differentially modulate neural processing of basic sensory events (i.e., syllables) and temporal chunks constructed based on knowledge (i.e., multisyllabic words). Recent human neurophysiological results have shown that cortical activity could concurrently follow linguistic units of different sizes, e.g., syllables and phrases (Ding et al., 2016). In this study, we used isochronous syllable sequences as the stimulus, in which neighboring syllables combined to form bisyllabic words (Fig. 1*A*). With such an isochro-

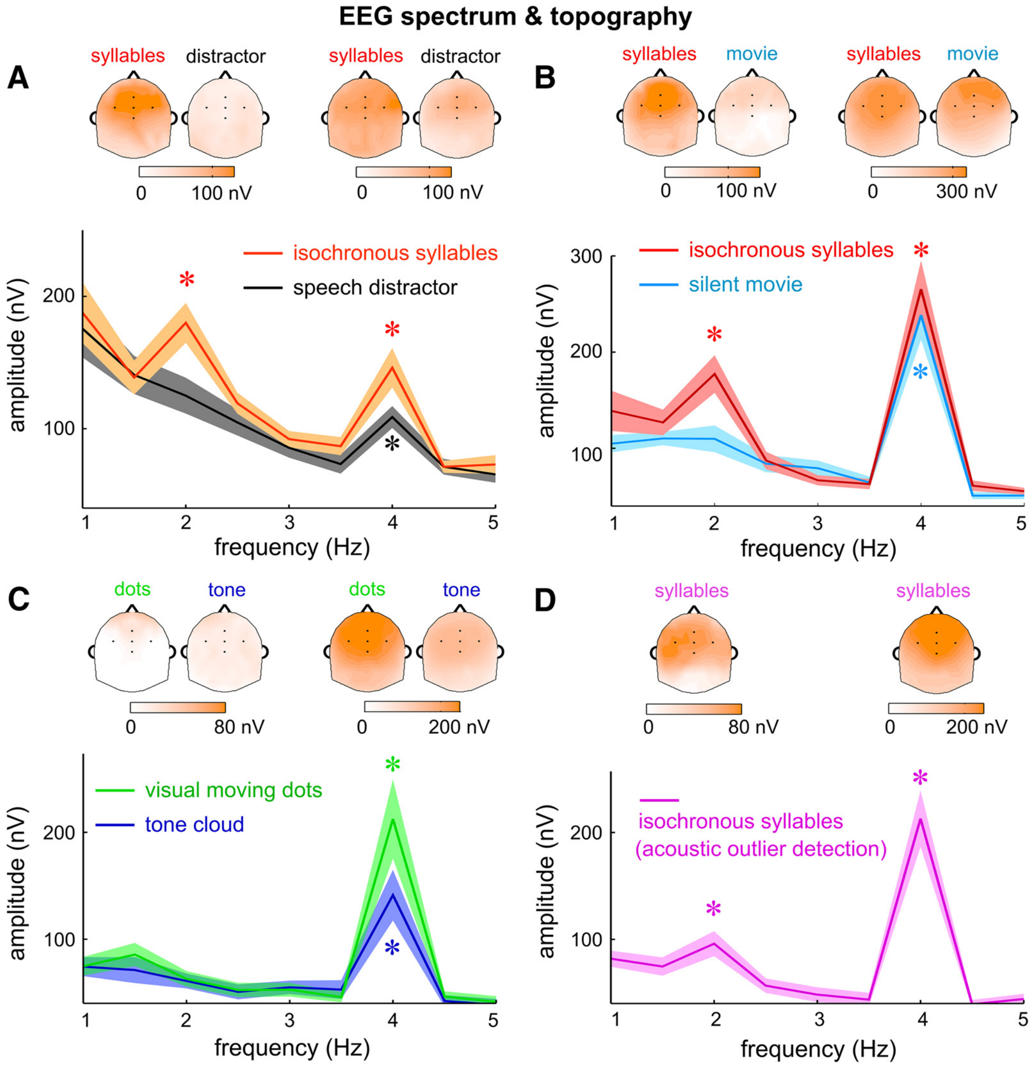


Figure 3. Attentional and task modulation of the neural response spectrum. The EEG spectrum is the root mean square over participants and channels. ***A–D***, Results from Experiments 1– 4 respectively. Different experimental conditions are color coded and the attentional focus of the participants in each condition is labeled. Stars indicate frequency bins with amplitude higher than the amplitude averaged over a 1-Hz-wide neighboring frequency region (*$p < 0.005$, bootstrap). Response peak at the syllable rate was observed in all tested conditions. Response peak at the word rate, however, was only observed when the isochronous syllable sequence was attended to. The topographic plots of the EEG response at the syllable and word rates are shown above the corresponding spectrum, which generally shows a central-frontal distribution. In the topographic plots, the EEG response is normalized by subtracting mean amplitude over a 1 Hz neighboring frequency region (excluding the target frequency). The five black dots in the topographic plots show the position of FCz (middle), Fz (upper), Cz (lower), FC3 (left), and FC4 (right). Figure 3-1 (available at https://doi.org/10.1523/JNEUROSCI.2606-17.2017.f3-1) shows additional spectral analysis of Experiment 3.

nous syllable sequence, neural responses tracking syllables (acoustic events) and bisyllabic words (temporal chunks) were separately tagged in frequency, allowing simultaneous monitoring of syllabic-level and word-level neural processing.

The participants' attentional focus and task were differentially manipulated in four experiments. Experiments 1–3 investigated whether the neural construction of multisyllabic spoken words was degraded when the listeners attended to a sensory distractor that ranged from a spoken passage, to a silent movie, to a basic auditory/visual stimulus with no linguistic content. These distractors all diverted attention but their processing pathways overlapped at different levels with the processing pathway for isochronous syllable sequences. By using different distractors, we could tease apart the influences of attention and competition in neural resources caused by overlapping processing pathways. Experiment 4, which did not present any sensory distractor, tested whether listeners could group syllables into words when they attended to basic auditory features rather than lexical information.

All experiments analyzed the steady-state neural response to an isochronous syllable sequence that alternated between word states and random states (Fig. 1*A*). In the word states, neighboring two syllables formed a bisyllabic word and in the random states syllables were presented in a random order. The EEG responses during the word states were analyzed. The response at the word rate (i.e., 2 Hz) was viewed as a neural marker for the grouping of syllables into bisyllabic words, while the response at the syllable rate (i.e., 4 Hz) was a neural marker for syllabic processing.

The first experiment used a dichotic listening paradigm and listeners were exposed to two concurrent speech streams, one to each ear (Fig. 1*B*). One speech stream was the isochronous syllable sequence while the other speech stream was a spoken passage that was time compressed, i.e., speeded up, by a factor of 2.5 to increase task difficulty. The experiment contained two blocks. In one block, the participants were asked to attend to the spoken passage and answered comprehension questions (cor-

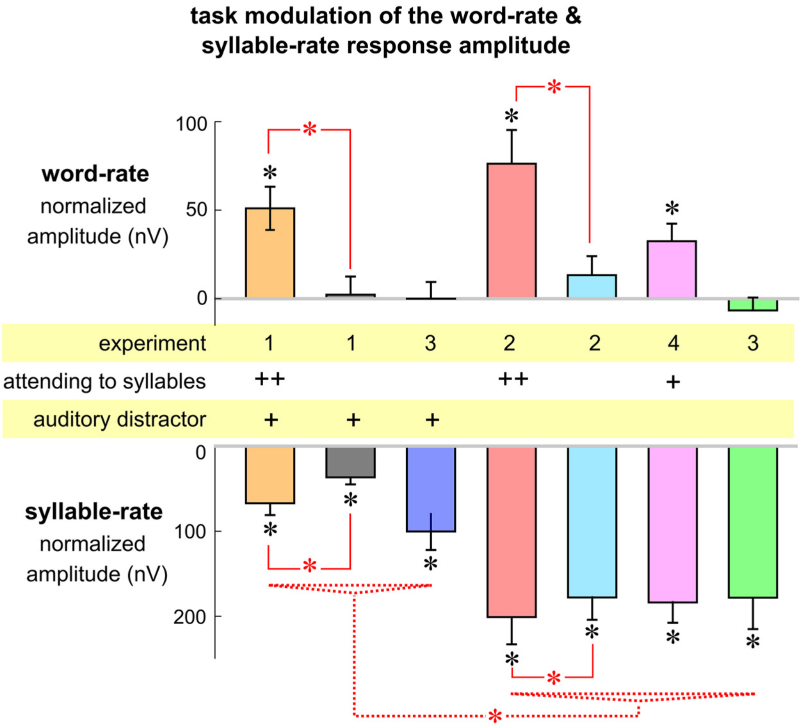


Figure 4. The influence of attention and tasks on the strength of word-rate and syllable-rate responses. Different experimental conditions are color coded the same way they are coded in Figure 3. The experiment each condition belongs to, whether the task is related to the isochronous syllables (++ for a word-level task and + for a speaker gender detection task), and whether the distractor is presented auditorily (+ for yes) are labeled in the middle of the figure. Response amplitude at each target frequency is normalized by subtracting the response amplitude averaged over a 1-Hz-wide neighboring frequency region (excluding the target frequency) to reduce the influence of background broadband neural activity. Black stars indicate that the normalized response amplitude is significantly >0 ($p < 0.001$, bootstrap, false discovery rate corrected). Red stars indicate significant differences between conditions ($p < 0.001$, bootstrap, false discovery rate corrected). Paired comparisons within the same experiment are also shown (solid red lines, star: $p < 0.001$, bootstrap). Unpaired comparisons across experiments were only applied to test whether the syllable-rate response is weaker in conditions with an auditory distractor than in conditions without an auditory distractor (dotted red lines). The star indicates a significant difference between any two conditions across the groups ($p <$ 0.001, bootstrap).

rect rate, 84 ± 2%, mean ± SEM over participants throughout this article). In the other block, they attended to the syllable sequence and had to indicate whether an additional word presented after the syllable sequence appeared in the sequence (correct rate, 77 ± 2%).

The EEG response spectrum averaged over participants and channels is shown in Figure 3*A*. When attention was directed to the isochronous syllable sequence, two peaks were observed in the EEG spectrum, one at the syllable rate ($p = 10^{-4}$, bootstrap) and the other at the word rate ($p = 10^{-4}$, bootstrap). A spectral peak was viewed as statistically significant if its amplitude was significantly stronger than the power averaged over a 1 Hz neighboring frequency region. The response topography showed that each response peak had a broad spatial distribution and was centered near channel FCz (Fig. 3*A*). When attention was directed to the speech distractor, however, a single response peak was observed at the syllable rate ($p = 10^{-4}$, bootstrap), while the neural response at the word rate was no longer significantly stronger than the response averaged over neighboring frequency bins ($p =$ 0.58, bootstrap). Attention to the isochronous syllables significantly increased the response amplitude at the word rate ($p =$ 0.02, bootstrap) and at the syllable rate ($p = 0.0005$, bootstrap). Nonetheless, the amplitude of the word-rate response was modulated by 95% while the amplitude of the syllable-rate response was only modulated by 46% (Fig. 4, 1 − the ratio between the gray and orange bars). These results demonstrated that selective attention had a much stronger influence on the neural representation of linguistically defined temporal chunks (i.e., words) than the neural representation of acoustic events (i.e., syllables).

In Experiment 1, neural processing of the speech distractor strongly overlapped with the neural processing of spoken words. Therefore, it was unclear whether top–down attention or a competition of other neural resources led to the strong modulation of the neural responses to words. To address this issue, Experiment 2 diverted top–down attention using a cross-modal distractor. In this experiment, the isochronous syllable sequence was identically presented to both ears and participants either listened to speech or watched a silent movie with subtitles. When the participants listened to speech, they indicated after each trial whether they heard more animate words or inanimate words (see Materials and Methods; correct rate, 81 ± 3%).

The EEG response spectrum in Experiment 2 is shown in Figure 3*B*. A word-rate response peak was observed when the participants attended to the isochronous syllables ($p = 10^{-4}$, bootstrap). A marginally significant word-rate response was observed when the participants watched a movie ($p = 0.098$, bootstrap). The attention-related change in response amplitude was 83% at the word rate but only 11% at the syllable rate (Fig. 4, 1 − the ratio between the cyan and red bars). These results showed that even without any competing auditory input, the word-level neural representation was still strongly modulated by attention.

Although a silent movie is a classic distractor used in passive listening experiments, it does not impose heavy processing load and cannot guarantee that the participants constantly focus on the movie. Additionally, the processing of subtitles may partly overlap with the processing of spoken words. Therefore, in Experiment 3, we engaged the participants in a challenging visual or auditory task that did not require any linguistic processing and tested whether a challenging sensory task (within-modal or cross-modal) was sufficient to block the grouping of syllables into words.

The auditory task in Experiment 3 relied on a dichotic listening paradigm. The auditory distractor was an isochronous tone sequence repeating at 3 Hz, embedded in a tone cloud (Fig. 1*D*). The participants had to detect frequency deviants in the tone sequence and report how many deviants occur in each trial ($N =$ 0, 1, or 2). In the visual task, participants saw cyan and orange dots moving on the screen and dots of a randomly selected color showing partly coherent motion. Occasionally, the coherent motion briefly reversed in direction and the participants had to report how many times such reversals occurred in each trial ($N = 0$, 1, or 2). The participants' response was correct in 81 ±

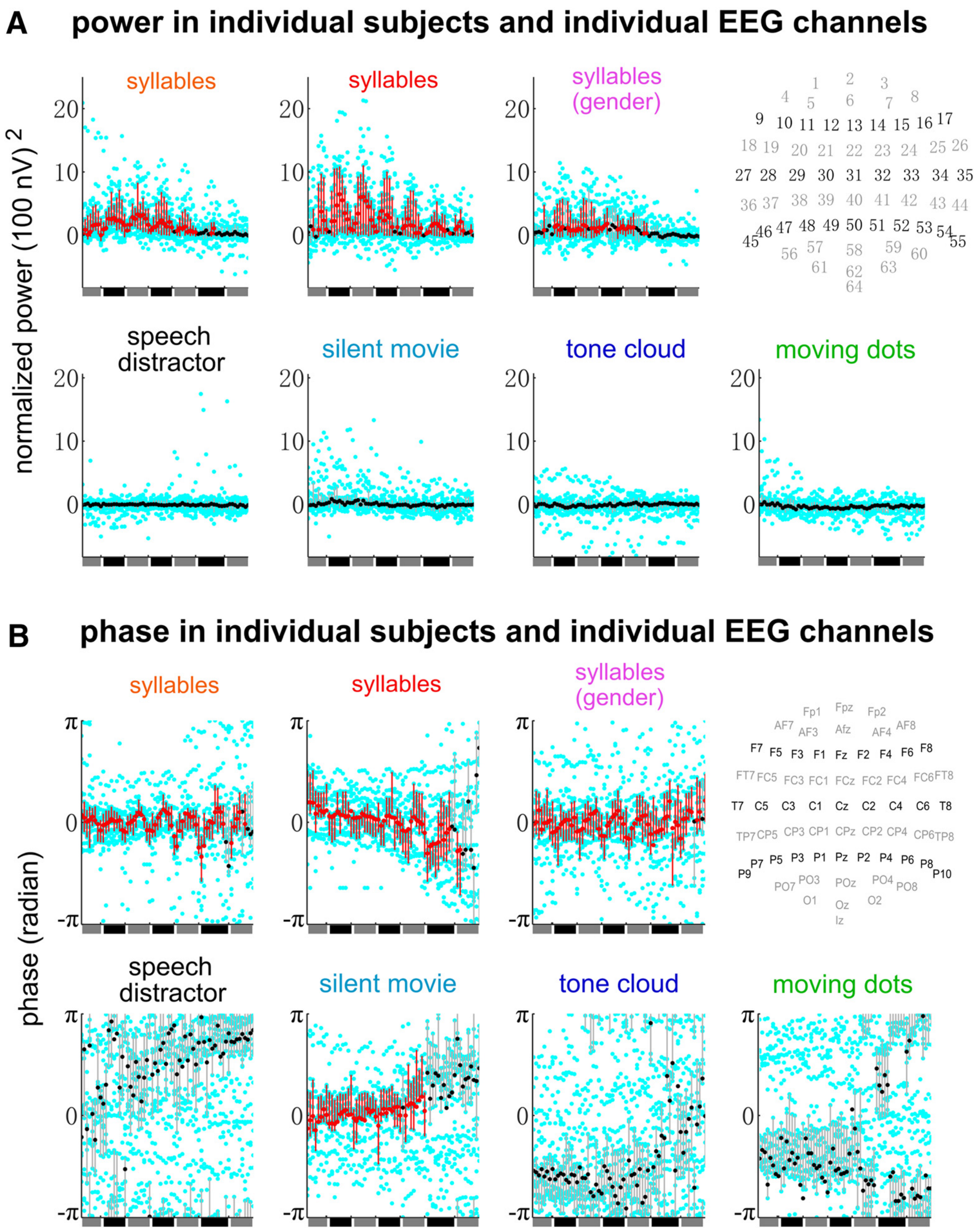


Figure 5. Word-rate response power and phase in individual EEG channels and individual participants. The power is the evoked power (i.e., power of the EEG waveforms averaged over trials) and is normalized by subtracting the power averaged over neighboring frequency bins (≤0.5 Hz on each side). Data from each participant are shown by cyan dots and the average over participants is shown by black dots. The *x*-axis represents EEG channels and the channel index, from 1 to 64, goes from left to right. The approximate scalp position of each channel is shown at the upper right corner of each panel. A vertical bar shows the range between the 25th and 75th percentile for power, and the minimal phase range covering 50% participants for the response phase. The bar is red if and only if the normalized power is significantly >0 in ***A*** or if the interparticipant phase coherence is significantly higher than chance in ***B*** ($p < 0.01$; see Materials and Methods; false discovery rate corrected). The experimental condition is labeled in the title of each plot and the 3 plots in the upper row of panels A and B show the conditions where the isochronous syllable sequence was attended to.

10% and 81 ± 12% trials for the auditory and visual tasks respectively.

The results of Experiment 3 are shown in Figure 3C. For both tasks, a significant response peak was observed at the syllable rate ($p = 10^{-4}$, bootstrap), but not at the word rate ($p > 0.5$, bootstrap). The syllable-rate response was weaker in the auditory task ($p = 0.0005$, bootstrap), probably due to the masking of the tone cloud. These results clearly demonstrated that knowledge-based grouping of syllables into words can be blocked when participants are engaged in demanding sensory tasks that do not involve language processing.

For the auditory task in Experiment 3, no response tracking the 3 Hz tone sequence was observed in the response spectrum in Figure 3C. The reason is that the EEG waveform was averaged based on the onset of each word state and therefore only revealed the response component phase-locked to the onset of the word states. If the response was averaged based on the onset of the 3 Hz tone sequence, a strong 3 Hz response was observed (Fig. 1–3).

Experiments 1 and 2 showed that neural activity can track the word rhythm when participants attended to the isochronous syllable sequence and performed a word-related task (judging whether a word appeared in the sequence in Experiment 1 and judging the animacy of words in Experiment 2). It remains unclear, however, whether attention to the right sensory input or the word-related task drives the grouping of syllables into words. This question is investigated in Experiment 4, in which the participants attended to the

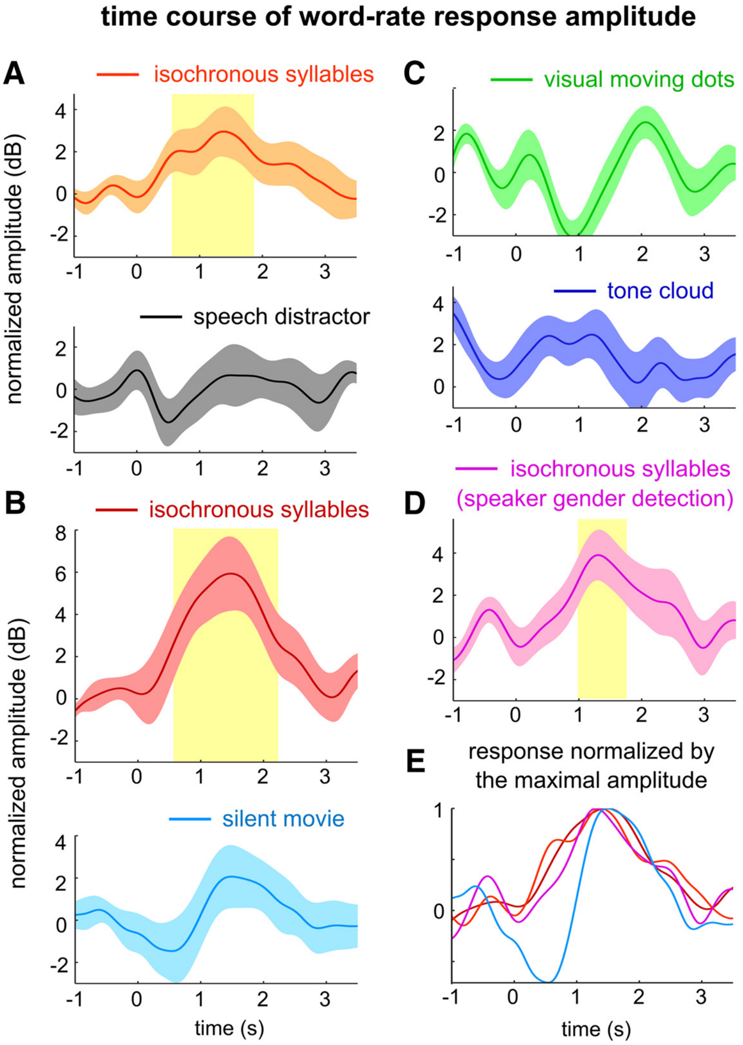


Figure 6. Temporal dynamics of the EEG response to words. ***A–D***, Instantaneous amplitude of the word-rate EEG responses in Experiments 1–4. Time 0 indicates the onset of a word state in the isochronous syllable sequence. The instantaneous amplitude is the magnitude of the Hilbert transform of the EEG responses filtered ~2 Hz (see Materials and Methods). The instantaneous amplitude was normalized by subtracting the instantaneous amplitude averaged over a 1 s prestimulus interval. Shaded yellow regions show time intervals when the response amplitude is significantly larger than the maximal value in the prestimulus interval (yellow: $p < 0.01$; light yellow: $p < 0.05$, bootstrap, false discovery rate corrected). The word-rate response peaks ~1–1.5 s after the first word appears, in conditions when the word-rate response exceeds the baseline. ***E***, Comparisons between the response buildup time course across conditions with a significant word-rate response. Each curve was normalized by its maximal amplitude. The word-rate response tends to build up more slowly in the movie-watching condition. However, no significant difference in response latency was observed.

isochronous syllable sequence but performed a nonlinguistic task, i.e., detecting occasional changes of the gender of the speaker. In this experiment, the behavioral correct rate was 92.5 ± 1%.

The results of Experiment 4 are shown in Figure 3*D*. A statistically significant word-rate response was observed ($p = 10^{-4}$, bootstrap), on top of the strong syllable-rate response ($p = 10^{-4}$, bootstrap). Therefore, the word-rate response remained statistically significant when the participants performed a low-level nonlinguistic task.

The results in Experiments 1–4 are summarized in Figure 4. To reduce the influence of background neural activity, the response amplitude was normalized by subtracting the mean response amplitude averaged over neighboring frequency bins (0.5 Hz on each side). On top of the comparisons reported separately for each experiment, there was a clear pattern showing that the syllable-rate response was weakened by the presence of auditory distractors ($p < 0.0003$ between any two conditions differing in the presence of auditory distractor, bootstrap, false discovery rate corrected; Fig. 4, dotted red lines). This result is consistent with previous findings showing that the response to the auditory input from one ear is weakened by the auditory input from the contralateral ear (Fujiki et al., 2002; Ding and Simon, 2012a).

Furthermore, the word-rate response was stronger in Experiment 4 than any of the conditions in Experiments 1–3 when the isochronous syllable sequence was not attended to ($p < 0.02$, bootstrap). The amplitude of the word-rate response, however, was smaller than that observed for the speech-listening condition in Experiment 2 ($p = 0.009$, bootstrap). Since the same physical stimulus was used in Experiment 4 and in the speech-listening condition in Experiment 2, a lexical-level task can indeed enhance neural tracking of words.

On top of the group analysis in Figures 3 and 4, the power and phase of the word-rate EEG response are shown in Figure 5 for individual participants and individual EEG channels. When the isochronous syllable sequence was attended to, the word-rate response power was stronger than the power averaged over neighboring frequency bins (Fig. 5*A*) and the response phase was consistent over participants (Fig. 5*B*). When the syllable sequence was not attended to, there was, in general, no significant increase in the word-rate power or interparticipant phase coherence. The movie-watching condition, however, was an exception that showed a consistent word-rate response phase across participants in several channels. A possible explanation is that movie watching is not a challenging task, so that the participants may occasionally shift their attention to speech and a weak neural response to words is captured by the phase coherence, a statistical measure more sensitive than response power (Ding and Simon, 2013).

The frequency-domain analyses in Figures 3–5 reveal steady-state properties of the neural tracking of syllables and words. The following analysis shows how the word-rate neural response evolves over time (Fig. 6*A–D*). During a word state, the word-rate response amplitude surpasses the response baseline (i.e., a 1 s period right before the onset of a word state) in conditions when the isochronous syllable sequence is attended to and in the movie-watching condition. For the conditions in which a significant word-rate response is observed, the response builds up following a similar time course across conditions (Fig. 6*E*) and stabilizes ~1 s after the word onset, similar to what is observed for neural tracking of phrases and sentences (Zhang and Ding, 2017). The peak latency and the latency when the response amplitude reaches half of the maximal amplitude (Fig. 6*E*, 0.5) are not significantly different across conditions ($p > 0.1$, bootstrap, not corrected for multiple comparisons).

Discussion

The current study investigated how attention and tasks may differentially modulate the neural tracking of acoustic events (i.e., syllables) and temporal chunks (i.e., words). In this study, the grouping of syllables into words relied only on top–down lexical knowledge (i.e., the mental dictionary), and not on bottom–up acoustic cues. We showed that attention to speech (Experiments 1–3), but not a lexical-meaning-related task (Experiment 4), is required to group syllables in that speech stream into multisyllabic words.

Neural processing of unattended auditory streams

The brain can detect statistical regularities in sounds even without top–down attentional modulation (Näätänen et al., 2007). For example, neural activity can entrain to intensity fluctuations in sound even when the listeners are not paying attention (Linden et al., 1987). Furthermore, when a random tone cloud turns into a fixed multitone sequence repeating in time, the brain can quickly detect such a transition even when attention is directed to other sensory stimuli (Barascud et al., 2016). The brain can also detect violations in multitone sequences that repeat in time (Sussman et al., 2007). Therefore, although attention can strongly modulate primitive auditory grouping (Carlyon et al., 2001; Shamma et al., 2011; Shinn-Cunningham et al., 2017), basic statistical regularities in sound can be extracted preattentatively.

Statistical regularities in sound can be extracted by bottom–up analysis of auditory features. In the current study, however, the grouping of syllables into words relied only on top–down knowledge, i.e., which syllables can possibly construct a valid multisyllabic word. Furthermore, the bisyllabic words were "hidden" in random syllables and the onset time was randomized. The word boundaries could only be determined by comparing the auditory input with word templates stored in the long-term memory. The current results showed that neural tracking of bisyllabic words was abolished when attention was directed to a sensory distractor that induced a high processing load. Therefore, although bottom–up grouping of basic auditory features into a sound stream may occur preattentatively, top–down schema-based grouping of syllables into words critically relies on attention.

Neural and behavioral studies of preattentative processing, however, always face two challenges. One challenge is that weak preattentative processing may fall below the sensitivity of the experimental measure. Although the current study found little evidence for preattentative grouping of syllables into words, it could not rule out the possibility that preattentative word grouping weakly occurs and is just not measurable by the current paradigm. What can be concluded, however, is that attention has a much larger effect on the neural grouping of syllables into words than it does on syllable encoding. The other challenge in studying preattentative processing is whether attention is fully controlled, i.e., whether the participants can split their attention or occasionally shift their attention to the stimulus they should ignore (Holender, 1986). In the current study, the weak word-rate response in the movie-watching condition was potentially caused by spontaneous shifts of attention.

Attention modulation of neural tracking of the speech envelope

Speech comprehension involves multiple processing stages, e.g., encoding acoustic speech features (Shamma, 2001), decoding phonemic information based on acoustic features (Mesgarani et al., 2014; Di Liberto et al., 2015), grouping syllables into words (Cutler, 2012), and grouping words into higher-level linguistic structures, such as phrases and sentences (Friederici, 2002). Previous studies have shown that neural tracking of the acoustic envelope, which corresponds to the syllable rhythm, is modulated by attention but remains observable for an unattended speech stream (Kerlin et al., 2010; Ding and Simon, 2012b; Power et al., 2012; Steinschneider et al., 2013; Zion Golumbic et al., 2013). This study, consistent with these previous studies, showed that the syllable-rate neural response, although reliably observed in all tested conditions, is modulated by top–down attention. Behaviorally, it has also been shown that a cross-modal sensory stimulus with no linguistic content could interfere with syllabic-level processing (Mattys and Wiget, 2011; Mitterer and Mattys, 2017).

Attention modulation of word processing

Previous neural and behavioral studies on preattentative processing of words mostly focused on semantic processing of words that have clear physical boundaries. Behavioral evidence suggests that cognitive processing of unattended spoken words is limited. Without paying attention, listeners cannot recall the spoken words they hear and cannot even notice a change in the language being played (Cherry, 1953). There is also evidence, however, for limited perceptual analysis of the unattended speech stream. For example, during dichotic listening, listeners can recall the gender of an unattended speaker (Cherry, 1953) and some listeners can notice their names in the unattended ear (Wood and Cowan, 1995; Conway et al., 2001). When shadowing words in the attended ear, performance can be influenced by semantically related materials presented in the unattended ear (Lewis, 1970; Treisman et al., 1974). These results suggest differential attentional modulation of different speech-processing stages. Without top–down attention, basic acoustic information such as the speaker's gender can be recalled and very salient words, such as one's name, are sometimes recalled. However, ordinary words cannot be recalled despite potential unconscious interference with processing of attended words.

It is found that the N400 response disappears for unattended auditory or visual words (Bentin et al., 1995; Nobre and McCarthy, 1995). On the other hand, visual experiments have shown that semantic processing can occur for words presented at the attended location even when these words are not consciously perceived (Luck et al., 1996; Naccache et al., 2002). Therefore, neural studies show that semantic processing of isolated words could be a subconscious process but requires attention. The current study extends these previous studies by showing that the phonological construction of words (i.e., the grouping of syllables into words) requires attention.

Here, the grouping of syllables was based purely on lexical knowledge. Natural speech, however, contains various prosodic cues for word boundaries that can facilitate lexical segmentation (Cutler, 2012). Future studies must investigate whether prosodic cues could trigger preattentative lexical analysis. Furthermore, in the current study, neural tracking of words was observed when the participants performed a gender-detection task that did not require lexical processing, in contrast to the N400 response to visual words, which was sensitive to the task (Chwilla et al., 1995). Gender detection, however, is an easy task and future studies are needed to investigate whether the response to words still exists for more difficult low-level tasks.

Findings in this study are largely consistent with the load theory for selective visual attention, which argues that in the presence of two sensory stimuli A and B, if stimulus A causes a high processing load it interferes with the processing of B, even when the processing of A and B has little overlap (Lavie, 1995). Support for the load theory mainly comes from visual experiments (Lavie, 2005), while the current study provides new cross-modal evidence for the theory: heavy processing load caused by visual motion detection can block auditory word recognition. In contrast, heavy processing load imposed by either a cross-modal distractor or a within-modal distractor cannot abolish neural encoding of syllables. Similarly, a recent study showed that during sleep, syllabic-level processing is largely preserved while word-level and syntactic-level processing is diminished (Makov et al., 2017).

Potential functions of low-frequency neural tracking of speech

The current data and previous studies (Steinhauer et al., 1999; Buiatti et al., 2009; Pallier et al., 2011; Peña and Melloni, 2012; Peelle et al., 2013; Ding et al., 2016, 2017; Farthouat et al., 2016; Meyer et al., 2016; Nelson et al., 2017) have shown that, during speech listening, cortical activity is concurrently synchronized to hierarchical linguistic units, including syllables, words, phrases, and sentences. Neural tracking of hierarchical linguistic units provides a plausible mechanism to map hierarchical linguistic units into coupled dynamic neural processes that allow interactions between different linguistic levels (van Wassenhove et al., 2005; Giraud and Poeppel, 2012; Goswami and Leong, 2013; Martin and Doumas, 2017).

Low-frequency neural tracking of sensory rhythms is a widely observed phenomenon. Neurophysiological evidence has shown that the phase of low-frequency neural oscillations can modulate neuronal firing (Lakatos et al., 2005; Canolty et al., 2006) and can serve as a mechanism for temporal attention and temporal prediction (Schroeder and Lakatos, 2009; Arnal and Giraud, 2012; Morillon et al., 2014). These neurophysiological hypotheses are naturally linked to the neurolinguistic hypothesis that when processing words, attention is directed to the word onsets (Astheimer and Sanders, 2009; Sanders et al., 2009). Furthermore, slow neural oscillations may also provide a neural context for the integration of faster neural activity falling into the same cycle of a slow neural oscillation (Buzsáki, 2010; Lisman and Jensen, 2013). Therefore low-frequency neural entrainment to temporal chunks may naturally provide a mechanism to put neural representations of sensory events into context and allow information integration across sensory events.

References

Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. Trends Cogn Sci 16:390–398. CrossRef Medline

Astheimer LB, Sanders LD (2009) Listeners modulate temporally selective attention during natural speech processing. Biol Psychol 80:23–34. CrossRef Medline

Barascud N, Pearce MT, Griffiths TD, Friston KJ, Chait M (2016) Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. Proc Natl Acad Sci U S A 113:E616–E625. CrossRef Medline

Bentin S, Kutas M, Hillyard SA (1995) Semantic processing and memory for attended and unattended words in dichotic listening: behavioral and electrophysiological evidence. J Exp Psychol Hum Percept Perform 21:54–67. CrossRef Medline

Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP (2013) Lexical influences on auditory streaming. Curr Biol 23:1585–1589. CrossRef Medline

Boersma P, Weenink D (2017) Praat: doing phonetics by computer. Available at http://www.praat.org/. Retrieved December 18, 2017.

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT.

Buiatti M, Peña M, Dehaene-Lambertz G (2009) Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. Neuroimage 44:509–519. CrossRef Medline

Buzsáki G (2010) Neural syntax: cell assemblies, synapsembles, and readers. Neuron 68:362–385. CrossRef Medline

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight RT (2006) High gamma power is phase-locked to theta oscillations in human neocortex. Science 313:1626–1628. CrossRef Medline

Carlyon RP, Cusack R, Foxton JM, Robertson IH (2001) Effects of attention and unilateral neglect on auditory stream segregation. J Exp Psychol Hum Percept Perform 27:115–127. CrossRef Medline

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979. CrossRef

Chwilla DJ, Brown CM, Hagoort P (1995) The N400 as a function of the level of processing. Psychophysiology 32:274–285. CrossRef Medline

Conway AR, Cowan N, Bunting MF (2001) The cocktail party phenomenon revisited: the importance of working memory capacity. Psychon Bull Rev 8:331–335. CrossRef Medline

Cutler A (2012) Native listening: language experience and the recognition of spoken words. Cambridge, MA: MIT.

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

Ding N, Simon JZ (2012a) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89. CrossRef Medline

Ding N, Simon JZ (2012b) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859. CrossRef Medline

Ding N, Simon JZ (2013) Power and phase properties of oscillatory neural responses in the presence of background activity. J Comput Neurosci 34:337–343. CrossRef Medline

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19: 158–164. CrossRef Medline

Ding N, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D (2017) Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front Hum Neurosci 11:481. CrossRef Medline

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Boca Raton, FL: CRC.

Farthouat J, Franco A, Mary A, Delpouve J, Wens V, Op de Beeck M, De Tiège X, Peigneux P (2017) Auditory magnetoencephalographic frequency-tagged responses mirror the ongoing segmentation processes underlying statistical learning. Brain Topogr 30:220–232. CrossRef Medline

Fodor JA (1983) The modularity of mind: an essay on faculty psychology. Cambridge, MA: MIT.

Friederici AD (2002) Towards a neural basis of auditory sentence processing. Trends Cogn Sci 6:78–84. CrossRef Medline

Fritz JB, Elhilali M, David SV, Shamma SA (2007) Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? Hear Res 229:186–203. CrossRef Medline

Fujiki N, Jousmaki V, Hari R (2002) Neuromagnetic responses to frequency-tagged sounds: a new method to follow inputs from each ear to the human auditory cortex during binaural hearing. J Neurosci 22:RC205. Medline

Gavornik JP, Bear MF (2014) Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat Neurosci 17:732–737. CrossRef Medline

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15: 511–517. CrossRef Medline

Goswami U, Leong V (2013) Speech rhythm and temporal structure: converging perspectives? Lab Phonol 4:67–92. CrossRef

Hannemann R, Obleser J, Eulitz C (2007) Top-down knowledge supports the retrieval of lexical information from degraded speech. Brain Res 1153: 134–143. CrossRef Medline

Holender D (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: a survey and appraisal. Behav Brain Sci 9:1–23. CrossRef

Jones JA, Freyman RL (2012) Effect of priming on energetic and informational masking in a same-different task. Ear Hear 33:124–133. CrossRef Medline

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". J Neurosci 30:620–628. CrossRef Medline

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J Neurophysiol 94:1904–1911. CrossRef Medline

Lashley KS (1951) the problem of serial order in behavior. In: Cerebral mechanisms in behavior, the Hixon symposium (Jeffress LA, ed), pp 112–136. New York: Wiley.

Lavie N (1995) Perceptual load as a necessary condition for selective attention. J Exp Psychol Hum Percept perform 21:451–468. CrossRef Medline

Lavie N (2005) Distracted and confused?: Selective attention under load. Trends Cogn Sci 9:75–82. CrossRef Medline

Lewis JL (1970) Semantic processing of unattended messages using dichotic listening. J Exp Psychol 85:225–228. CrossRef Medline

Linden RD, Picton TW, Hamel G, Campbell KB (1987) Human auditory steady-state evoked potentials during selective attention. Electroencephalogr Clin Neurophysiol 66:145–159. CrossRef Medline

Lisman JE, Jensen O (2013) The theta-gamma neural code. Neuron 77:1002–1016. CrossRef Medline

Lu K, Xu Y, Yin P, Oxenham AJ, Fritz JB, Shamma SA (2017) Temporal coherence structure rapidly shapes neuronal interactions. Nat Commun 8:13900. CrossRef Medline

Luck SJ, Vogel EK, Shapiro KL (1996) Word meanings can be accessed but not reported during the attentional blink. Nature 383:616–618. CrossRef Medline

Makov S, Sharon O, Ding N, Ben-Shachar M, Nir Y, Zion Golumbic E (2017) Sleep disrupts high-level speech parsing despite significant basic auditory processing. J Neurosci 37:7772–7781. CrossRef Medline

Martin AE, Doumas LA (2017) A mechanism for the cortical computation of hierarchical linguistic structure. PLoS Biol 15:e2000663. CrossRef Medline

Mattys SL, Wiget L (2011) Effects of cognitive load on speech recognition. J Mem Lang 65:145–160. CrossRef

Mattys SL, Brooks J, Cooke M (2009) Recognizing speech under a processing load: dissociating energetic from informational factors. Cogn Psychol 59:203–243. CrossRef Medline

McDermott JH, Wrobleski D, Oxenham AJ (2011) Recovering sound sources from embedded repetition. Proc Natl Acad Sci U S A 108:1188–1193. CrossRef Medline

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

Meyer L, Henry MJ, Gaston P, Schmuck N, Friederici AD (2017) Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb Cortex 27:4293–4302. CrossRef Medline

Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48:139–148. CrossRef Medline

Mitterer H, Mattys SL (2017) How does cognitive load influence speech perception? An encoding hypothesis. Atten Percept Psychophys 79:344–351. CrossRef Medline

Morillon B, Schroeder CE, Wyart V (2014) Motor contributions to the temporal precision of auditory attention. Nat Commun 5:5255. CrossRef Medline

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118:2544–2590. CrossRef Medline

Naccache L, Blandin E, Dehaene S (2002) Unconscious masked priming depends on temporal attention. Psychol Sci 13:416–424. CrossRef Medline

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen L, Koopman H, Cash SS, Naccache L, Hale JT, Pallier C, Dehaene S (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci U S A 114:E3669–E3678. CrossRef Medline

Nobre AC, McCarthy G (1995) Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. J Neurosci 15:1090–1098. Medline

Pallier C, Devauchelle AD, Dehaene S (2011) Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci U S A 108:2522–2527. CrossRef Medline

Patel AD (2008) Music, language, and the brain. New York: Oxford UP.

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

Peña M, Melloni L (2012) Brain oscillations during spoken sentence processing. J Cogn Neurosci 24:1149–1164. CrossRef Medline

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech? Eur J Neurosci 35:1497–1503. CrossRef Medline

Sanders LD, Ameral V, Sayles K (2009) Event-related potentials index segmentation of nonsense sounds. Neuropsychologia 47:1183–1186. CrossRef Medline

Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32:9–18. CrossRef Medline

Shamma S (2001) On the role of space and time in auditory processing. Trends Cogn Sci 5:340–348. CrossRef Medline

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34:114–123. CrossRef Medline

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182–186. CrossRef Medline

Shinn-Cunningham B, Best V, Lee AK (2017) Auditory object formation and selection. In: The auditory system at the cocktail party (Middlebrooks JC, Simon JZ, Popper AN, Fay RR, eds), pp 7–40. Heidelberg: Springer International Publishing.

Snyder JS, Alain C, Picton TW (2006) Effects of attention on neuroelectric correlates of auditory stream segregation. J Cogn Neurosci 18:1–13. CrossRef Medline

Steinhauer K, Alter K, Friederici AD (1999) Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nat Neurosci 2:191–196. CrossRef Medline

Steinschneider M, Nourski KV, Fishman YI (2013) Representation of speech in human auditory cortex: Is it special? Hear Res 305:57–73. CrossRef Medline

Sussman ES, Horváth J, Winkler I, Orr M (2007) The role of attention in the formation of auditory streams. Percept Psychophys 69:136–152. CrossRef Medline

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136. CrossRef Medline

Treisman A, Squire R, Green J (1974) Semantic processing in dichotic listening? A replication. Mem Cogn 2:641–646. CrossRef Medline

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102:1181–1186. CrossRef Medline

Wood N, Cowan N (1995) The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? J Exp Psychol Learn Mem Cogn 21:255–260. CrossRef Medline

Woods KJ, McDermott JH (2015) Attentive tracking of sound sources. Curr Biol 25:2238–2246. CrossRef Medline

Zhang W, Ding N (2017) Time-domain analysis of neural tracking of hierarchical linguistic structures. Neuroimage 146:333–340. CrossRef Medline

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980–991. CrossRef Medline