The Role of Rhythm in Auditory Speech Comprehension

Authors: Liangjie Chen, Liu Lei, Ge Zhongshu, Yang Xiaodong, Li Liang, Li Liang

Date: 2021-12-29T22:56:37+00:00

Abstract

Speech comprehension is the psychological process by which a listener receives external speech input and acquires meaning. In daily communication, auditory speech comprehension is influenced by multi-scale rhythmic information, commonly including three aspects of external rhythms: prosodic structure rhythm, contextual rhythm, and speaker' s body language rhythm. These alter processes such as phoneme discrimination, lexical perception, and speech intelligibility in listeners during speech comprehension. Internal rhythms manifest as neural oscillations in the brain, which can encode hierarchical features of external speech input at different temporal scales. Neural entrainment between external rhythmic stimuli and internal neural activity can optimize the brain' s processing of speech stimuli, and is further modulated by the listener' s top-down cognitive processes to enhance the internal representation of target speech. We propose that this may be a key mechanism for realizing the interconnection between internal and external rhythms and their joint influence on speech comprehension. Elucidating the mechanisms of internal and external rhythms and their interconnections can provide a research window for understanding speech as a complex sequence with structural regularities at multiple hierarchical temporal scales.

Full Text

The Role of Rhythm in Auditory Speech Understanding

Authors: Chen Liangjie, Liu Lei, Ge Zhongshu, Yang Xiaodong, Li Liang (Corresponding Author)
Affiliation: School of Psychological and Cognitive Sciences, Peking University, Beijing, China
Corresponding Author Email: liliang@pku.edu.cn

Abstract

Speech understanding is a psychological process through which listeners receive external speech input and extract meaning. In daily communication, auditory speech comprehension is influenced by multi-scale rhythmic information, primarily manifesting in three types of external rhythms: prosodic structure rhythm, contextual rhythm, and speaker body language rhythm. These rhythms alter phoneme discrimination, lexical perception, and speech intelligibility during speech understanding. Internal rhythms are represented by neural oscillations in the brain, which encode hierarchical features of external speech input across different temporal scales. Neural entrainment between external rhythmic stimuli and internal neural activity can optimize the brain's processing of speech signals and further enhance the internal representation of target speech through top-down modulation of the listener' s cognitive processes. We propose that neural entrainment may be the key mechanism linking internal and external rhythms to jointly influence speech understanding. Elucidating these mechanisms and their connections provides a research window for understanding speech as a complex sequence with structural regularities across multiple hierarchical temporal scales.

Keywords: rhythm, speech understanding, neural oscillation, neural entrainment, top-down modulation

1. Introduction

From the cycles of life and death to the alternation of day and night, the natural world is filled with complex rhythmic variations. Activities such as drumming, dancing, or singing involve clapping, stepping, or vocalizing that typically follow certain periodic cycles. Rhythm is ubiquitous throughout human cultural evolution, carrying important functions for information transmission and serving as a crucial medium for social communication and interaction. For a long time, rhythm research has primarily focused on sensory processing, neglecting its role in more complex speech understanding. Only recently have researchers begun to systematically examine how rhythm influences speech understanding and to reveal its underlying mechanisms by recording listeners' internal neural activity.

The inherent temporal organization of rhythm regulates how individuals communicate and interact. As an important channel of information exchange in human society, spoken language possesses rich rhythmic characteristics. There are two approaches to determining whether an object has rhythmic properties: one emphasizes temporal regularity, while the other emphasizes structural relationships in time. The former defines rhythm as coordinated or periodic rhythm, meaning the constant repetition of fixed intervals and patterns. Examples include the "tick-tock" sounds of clock rotation and the rhythmic beating of a healthy heart, both exhibiting temporal regularity or near-regularity. Speech rhythm, however, aligns more with the latter definition: a stable relationship between given attributes or combinations of attributes over a time span. For

instance, knots in a tree trunk cause sawing to stall, yet we still perceive the lumberjack' s back-and-forth movements as rhythmic.

In the field of linguistics, early research on the perception of "machine-gun"Spanish, "Morse code" English, and Japanese pronunciation led researchers to focus on isochrony phenomena across different language families, classifying rhythm perception into stress-timed, syllable-timed, and mora-timed forms. However, this classification overemphasizes inter-unit isochrony and lacks empirical support for "isochrony theory" in phonetic signal analyses across multiple language families. Later classification methods based on vowel duration variation proved more empirical, attempting to establish a more universal rhythmic quantification approach according to differences in the proportion of vowel intervals in spoken language. For example, stress-timed languages show more variable vowel durations compared to syllable-timed languages. These classification methods demonstrate that speech, unlike activities formed by a single oscillator with specific interval repetitions, lacks objective isochronous periodicity, yet can still be intuitively perceived as rhythmic activity. Similar to rhythm in music, individual attributes in speech such as pitch variation or syllable duration can also produce subjective rhythmic sensations. However, focusing solely on individual attribute metrics cannot encompass all features of speech rhythm, which also depends on factors including overall loudness changes and speech rate. These factors collectively act on the listener' s perceptual processing, enabling the perception of rhythm in speech.

Speech understanding is a psychological process in which listeners acquire meaning from external speech input (such as target speech) and background information (such as context or non-verbal information), involving processing at different hierarchical levels including syllables, words, and sentences. Spoken language exhibits certain rhythmic characteristics in its prosodic structure, such as stress placement and speech rate, which influence listeners' comprehension of target speech. When a speaker' s articulation rate or syllable production rate exceeds the normal range (3-8 Hz), speech intelligibility decreases significantly. In contrast, background contextual rhythm alters listeners' syllable-level perception; for example, presenting a regular pure-tone sequence beforehand changes how individuals perceive subsequent consonants, with faster rhythmic sequences causing listeners to perceive consonants more as /w/ than /b/. Background information is not limited to acoustic-level changes. Since the temporal envelope of speech, the vocal tract activity of the speaker, and body movements are highly correlated, speech understanding is also influenced by non-verbal rhythms such as speaker body language, which includes facial movements, postures, and gestures. A speaker' s facial movements often share similar rhythmic characteristics with the temporal envelope of speech, helping listeners better understand speech information. Accordingly, this paper defines external rhythms as physical inputs with rhythmic features in the objective world that can influence speech understanding during auditory speech comprehension. We will discuss how three common external rhythms—prosodic structure rhythm, contextual rhythm, and speaker body language rhythm—affect speech understanding at

three levels: phonemes, words, and sentences, thereby illustrating the role of external rhythms in speech comprehension.

How does the listener' s brain utilize external rhythms to facilitate or alter speech understanding? This process is believed to be closely linked to internal rhythms, namely neural oscillations—rhythmic, synchronized electrical activities of intracranial neuronal populations. Neural oscillations, as cortical neuronal ensemble activities, are thought to mediate various cognitive processing stages, including speech processing and interference suppression. Recent research suggests that internal rhythmic activity may be influenced by external rhythms, exhibiting a phenomenon where internal and external rhythms converge over time, known as neural entrainment. When internal rhythms entrain to external target speech, listeners demonstrate better speech comprehension performance. Additionally, various high-level cognitive processes in speech understanding can modulate neural entrainment, such as selective attention, prior grammatical knowledge, and contextual expectation.

Based on this, we propose that neural entrainment may be the key mechanism for linking and coordinating internal and external rhythms during speech understanding. This paper first discusses how three common external rhythms influence auditory speech understanding, demonstrating the universality of rhythm' s impact on speech comprehension. We then summarize the functions of neural oscillations as internal rhythms in speech understanding. Finally, by examining the role of neural entrainment in speech processing and its modulation by top-down cognitive processes, we discuss the possibility that neural entrainment serves as a mechanism linking internal and external rhythms. Future research should investigate the significance of rhythm in auditory speech understanding across different hierarchical levels, scales, and contexts.

2. External Rhythms and Speech Understanding

Speech production unfolds over time, making temporal regularity crucial for listeners to understand information. To comprehend speech content, listeners must perceive the temporal organization of phonemes, syllables, words, and phrases from continuous speech streams based on external rhythmic features. In this section, we categorize external rhythms that influence speech understanding into three common types based on speech input and background information: prosodic structure rhythm, contextual rhythm, and speaker body language rhythm.

2.1 Prosodic Structure Rhythm Alters Sentence Intelligibility

Prosodic structure rhythm manifests differently in reading and spoken communication. In visual reading, the dynamic combination of words with different syllable counts influences local phrase analysis and overall sentence integration. However, reading, which relies primarily on visual input, cannot directly provide prosodic structure information and requires readers to rely on internal

representations such as subvocalization. This section focuses on prosodic structure rhythm in auditory scenes, specifically features such as syllable duration, inter-syllable intervals, and stress distribution in spoken language.

Inter-syllable intervals directly affect speech intelligibility. Researchers compressed sentences temporally to reduce inter-syllable pauses, resulting in faster overall speech rate and dramatic decreases in listeners' sentence intelligibility. Listeners struggled to process speech stimuli with disrupted prosodic structure rhythm, but this could stem from either sentence processing dependence on specific rhythmic sensory input or destroyed intra-syllable acoustic structures that hinder recognition after temporal compression. To address this question, researchers segmented compressed speech waveforms at equal intervals, leaving syllables within each segment still compressed, then inserted silent intervals after each fragment to create artificial rhythmic properties. Intelligibility for such sentences was restored. Notably, intelligibility only recovered when inserted intervals followed a regular pattern; irregular intervals had no effect. Therefore, speech understanding depends on speech' s intrinsic rhythmic properties. The process of inserting silent intervals into compressed sentences can be understood as "repackaging" syllables within the sentence—segmenting the temporal waveform into different parts. These packages are transmitted to both ears at a specified rate, helping listeners predict the maximum information transmission rate within each package, thereby restoring speech intelligibility to some extent.

The influence of prosodic structure rhythm on intelligibility reflects the auditory system's adaptability when processing information streams at different transmission rates. In natural speech, syllable pause duration involves two main factors: the biomechanical properties of human articulatory organs and the neurodynamic properties of the brain. Intrinsic oscillations of articulatory organs/brain modulate lip movements and speech temporal envelopes at approximately 5 Hz, thereby regulating silence duration. The second factor is speech' s hierarchical prosodic structure. For example, when a syllable occurs within a word, the subsequent pause is typically short, but when it coincides with a boundary of a higher-level linguistic structure (such as a prosodic word, prosodic phrase, or intonation phrase), silence gradually lengthens. Silences in speech provide the brain with additional time to process preceding syllables. When silence duration is reduced or expanded, violating natural language temporal regularities, it increases listeners' processing load and consequently disrupts sentence intelligibility.

Beyond pause duration, pause location also alters listeners' perception of speech rhythm, primarily involving prosodic boundaries in spoken sentences. These boundaries are associated with perceived pauses, pre-boundary syllable lengthening, and phrase-final pitch, so prosodic boundary perception facilitates listeners' segmentation of speech into different hierarchical chunks and is closely related to perceived fluency and intelligibility. As a tonal language, Chinese shows prosodic boundary effects in structural analysis, semantic processing, and emotional perception of spoken language. Recent research using Chinese ambigu-

ous phrases that can be interpreted as either modifier-noun constructions or verb-object constructions found that when listeners attend to prosodic information, prosodic boundaries alter their structural analysis of ambiguous phrases. Prosodic boundaries help listeners analyze sentence structure in ambiguous contexts by eliminating structural ambiguity and promoting speech intelligibility. Additionally, stress placement in Chinese prosody can alter listeners' selective attention to different lexical positions in speech, with stronger processing of post-stress words.

2.2 Contextual Rhythm Alters Lexical and Phonemic Perception

The acoustic scene preceding and following target speech is generally referred to as context, which may be temporally adjacent or non-adjacent to the target speech. Context influences speech understanding primarily through speech rate because listeners depend on relative rate cues provided by the context for lexical perception or boundary segmentation. Speakers modulate speech rate by adjusting intervals between vowels and consonants, and the distribution of these intervals reflects rhythmic properties in utterances. When speakers talk slowly, listeners tend to omit function words (such as "or" or "are") from sentences, exhibiting a word disappearance phenomenon at the perceptual level. Interestingly, increasing speech rate causes listeners to perceive function words that were not actually present. This effect intensifies with longer contextual duration. Baese-Berk et al. (2014) manipulated both global context (the entire material) and distal context (the sentence containing the target word) speech rates, finding that the influence of global speech rate on target word quantity perception increased over time, with perceived word count decreasing as speech rate slowed. These findings indicate that as contextual rhythm accelerates or decelerates, listeners' perception of word quantity in speech shifts in a compensatory direction to ensure stable perception, spontaneously adjusting subjective perception of subsequent word duration or boundary positions to match overall contextual rhythm. Notably, this phenomenon may be context-specific; artificially reducing speech intelligibility or using other tonal sequences does not affect listeners' word count identification.

Speech rate not only changes listeners' judgments of word quantity in context but also affects identification of vowels and consonants within words, as speech perception largely depends on recovering phonetic cues from specific frequency information. For example, a fast speech environment makes listeners more likely to judge an ambiguous vowel as a long vowel (e.g., /a, a:/) because syllable duration in adjacent context alters subjective evaluation of subsequent syllable duration; fast rhythmic contexts make listeners' objective time judgments shorter, making subsequent vowels sound relatively longer. The effect of contextual rate on phoneme boundaries is called phonetic boundary shift (PBS). This phenomenon also occurs in consonant perception, where faster rhythmic contexts increase the likelihood of perceiving ambiguous /ba/-/wa/ syllables as /wa/. Phoneme perception in speech is influenced by external rhythms from context.

Since this phenomenon can also be evoked in non-speech environments, such as pure-tone sequences, this rate-dependent perception is believed to involve general auditory processes.

Perception is never objective registration of sensory information. Like any form of perception, speech perception is context-relative and changes according to prior experience and background. The above studies demonstrate that under external rhythmic induction, listeners' perception of word quantity and syllable discrimination in given contexts changes. These results help explain why speech recognition ability declines under speech signal distortion conditions.

2.3 The Influence of Body Language Rhythm on Speech Understanding

Body language is a non-verbal communication mode through which speakers use facial activities and hand gestures to assist information expression. In face-to-face communication, body activities and speech rhythms perceived simultaneously by listeners often match at specific frequencies, facilitating their coupling. Just as an animated speaker gesturing with hands and feet more easily captures audience attention, these synchronized movements enhance speech comprehension.

A series of coordinated movements in speakers' articulatory organs manifest as cycles of vocal tract opening and narrowing. For example, producing /b/ requires closing the front portion of the vocal tract, creating a coordinated process between lip and jaw movements to achieve complete closure. Many studies have focused on the interaction between sound and movement in speech. When listeners observe speakers' lip movements, artificially altering movement rate affects their judgment of actual speech rate. In multi-talker scenarios, researchers have found that speakers' lip movement information can improve listeners' target speech identification performance. Beyond utilizing speakers' lip movements, listeners also use speakers' spontaneous hand movements to understand speech. Speakers often use gesture movements to highlight stress positions, and researchers have found that speakers' biphasic hand movements (up-and-down arm swinging) significantly alter listeners' perception of stress location in words.

Listeners can utilize non-acoustic body language rhythmic information to facilitate speech understanding, suggesting that shared prior knowledge may exist between listeners and speakers. The motor theory of speech perception proposes that speakers and listeners share a similar set of neuromotor commands. When listeners process speakers' movement information and map it onto their own commands, this facilitates understanding of speech content.

In summary, external rhythms influence auditory speech understanding across broad auditory and non-auditory stimuli. Context speech rate can alter listeners' subsequent phoneme discrimination and word quantity estimation. Intrinsic speech rhythm can change sentence intelligibility. Body language rhythm can

alter stress position perception. However, how our brain utilizes this rhythmic information to guide speech perception will be discussed next from the perspective of rhythmic oscillations in neuronal populations.

3. Neural Mechanisms of External Rhythm' s Influence on Speech Understanding

Early research on brain internal processes of auditory speech understanding primarily used event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI). Syllable detection and speech understanding involve ERP components such as N1-P2, N400, and P600. With methodological improvements, spontaneous neural oscillations recorded through intracranial electrodes and time-frequency analyses have become focal points, with increasing studies revealing speech understanding from the neural oscillation perspective. This section examines internal brain rhythm changes during speech processing and the role of neural entrainment phenomena.

3.1 Listeners' Internal Rhythms—Neural Oscillations

How are speech's intrinsic rhythmic characteristics represented in the brain, and how do external rhythms influence speech perception? To address these questions, researchers have begun investigating the role of internal brain rhythmic activity. Early on, electrical activity recorded from the scalp was considered background noise, but researchers later realized that oscillatory activity of neuronal populations reflects periodic changes in neuronal excitability. For instance, the instantaneous phase of oscillations reflects the excitability level of neuronal ensembles at a given moment. When the excitability phase of oscillations is adjusted so that high excitability of neuronal populations aligns with task-relevant sensory input, aligned input receives optimal processing. Therefore, internal brain rhythms may represent ideal neural oscillations for processing external rhythmic stimuli.

Neural oscillations are commonly divided by frequency into delta (1-4 Hz), theta (4-10 Hz), alpha (8-15 Hz), beta (12-30 Hz), and gamma (30-200 Hz) bands. In auditory speech processing, theta-band oscillations are believed to segment continuous speech signals into discrete word units, while delta-band oscillations combine segmented words into higher-level grammatical or semantic structures. Recent research on Chinese prosodic context processing also found that prosodic rhythm may facilitate speech comprehension by enhancing activity in frequency bands related to speech processing. Compared to irregular prosodic contexts, regular prosodic contexts elicit enhanced beta-band activity before target nouns and alpha-band activity after target nouns. Higher-frequency gamma-band envelope changes have been found to represent multi-level coding of speech in the power spectrum and are influenced by listeners' target selection.

Similar to hierarchical structures in speech, different frequency neural oscillations tend to couple in a hierarchical pattern. Low-frequency oscillations

(e.g., theta band) in the brain may reflect syllable-level processing, while high-frequency oscillations (e.g., gamma band) represent phoneme or articulatory feature information. Inter-frequency coupling reflects long-distance brain region communication and coordinates global neural network information integration. In primary auditory cortex (A1), gamma-band amplitude systematically varies with theta oscillation phase, and theta amplitude also couples with delta (1-2 Hz) phase. Interestingly, such effects are influenced by speech intelligibility. Compared to time-reversed speech (unintelligible), processing natural speech (intelligible) shows that listeners' left inferior frontal delta band and precentral theta band can modulate phase activity of 25 Hz oscillations in left auditory regions. Thus, different rhythmic neural oscillations can represent hierarchical information at different temporal scales in speech stimuli, integrating this information through mutual coordination to accomplish auditory speech understanding.

3.2 Neural Entrainment Connects Internal and External Rhythms

When external rhythmic stimuli are presented, listeners' brains exhibit phase alignment with external rhythms or increased power in the same frequency band. These processes are believed to result from the "resetting"of ongoing neural oscillation phases by external rhythmic stimuli. We refer to this temporal alignment between internal and external rhythms as neural entrainment. Researchers generally believe neural entrainment occurs based on the brain' s intrinsic rhythmic activity, which can maintain activity without continuous external input, allowing entrainment to persist for some time after external stimulus disappearance.

Common neural entrainment calculation methods include phase coherence between external stimuli and brain activity, and regression models connecting brain and stimuli in forward (e.g., temporal response functions) or backward (e.g., stimulus reconstruction) directions. Therefore, neural entrainment is sometimes called synchronization, or speech tracking when external stimuli are auditory speech.

In speech understanding, external rhythms may derive from syllables, word boundaries, or other acoustic cues. Neural entrainment can accomplish speech analysis through these external rhythmic features, extracting discrete linguistic components from continuous sound signals. The classical neural entrainment view holds that entrained neural activity phases align with prosodic or syllable boundaries in speech, such as tracking prosodic cues through delta oscillations and reflecting syllable and lexical structure through theta oscillations. Luo and Poeppel (2007) recorded magnetoencephalography signals while listeners processed natural speech and found that brain theta oscillation phase patterns stably followed syllable rhythms in spoken sentences. This study also found that when speech was embedded in noise, loss of external speech rhythmicity and decreased sentence intelligibility disrupted listeners' neural entrainment responses. Beyond interference from other acoustic stimuli, speech rate itself affects neural entrainment—once speech becomes too fast, listeners struggle to follow sentence

content and entrainment is interrupted. Interestingly, entrainment to physical acoustic features within speech is automatic, observable even during sleep. However, entrainment to linguistic units within sentences requires speech to be attended to or understood. Although studies on intelligibility and speech rate reflect neural entrainment' s role in speech processing, they still have limitations. Reducing speech intelligibility typically involves acoustic changes to stimuli, so observed differences in speech tracking responses may relate to altered sound input. Future research exploring the relationship between speech understanding and neural entrainment must carefully control the acoustic characteristics of speech stimuli.

In face-to-face conversation, listeners' speech understanding is influenced by body language. Park et al. (2016) found this process also involves entrainment between speaker movement and listener neural activity. By calculating the temporal pattern of speaker mouth opening area and listeners' neural activity in primary audiovisual cortex and left motor regions, they found significant entrainment effects at 1 Hz, with target speech comprehension predictable from entrainment synchrony. Listeners'brain entrainment to body speech enables real-time utilization of a "simulated" vocalization process in motor regions to help auditory regions predict upcoming sensory stimuli, thereby facilitating speech understanding.

Neural entrainment reflects how rhythmic information influences speech understanding. When the temporal regularity of speech as an external rhythm is perceived by listeners, resetting of internal neural activity makes internal rhythm patterns similar to external rhythms, making neural activity under similar phase patterns an ideal environment for speech understanding. However, neural entrainment is not merely passive following of external rhythmic information; it is also influenced by listeners' subjective regulation. We will further explore neural entrainment' s role in speech understanding by examining its modulation by several top-down cognitive processes involved in speech comprehension.

4. Top-Down Modulation of Neural Entrainment in Speech Understanding

Neural entrainment can dynamically select or enhance synchronization with external input according to listeners' current cognitive states, facilitating more targeted prediction of target information. During speech understanding, top-down modulation may derive from cognitive processes such as selective attention, prior grammatical knowledge, and expectations generated from speech context.

Noisy acoustic environments make target speech understanding difficult. Selective attention helps amplify entrainment differences between attended and unattended stimulus streams. Phase synchronization between attended stimuli and neural activity facilitates greater processing resources, while unattended stimuli are transmitted to non-optimal phase stages, making them easier to suppress, thereby aiding speech understanding in noisy environments. In multi-talker sce-

narios, when listeners selectively attend to a single speaker' s speech content, both auditory cortical regions (e.g., superior temporal gyrus) and higher-level brain regions (e.g., inferior frontal cortex, anterior temporal regions) show enhanced amplitude modulation of neural oscillations, with higher-level cortical regions exhibiting more pronounced selective enhancement of entrainment to attended speech. Furthermore, selective attention facilitates body language' s promotion of auditory speech understanding. When listeners pay more attention to speakers' lip movements, entrainment between left motor cortex and lip movements strengthens, directly predicting speech comprehension accuracy. Thus, neural entrainment across different brain regions can establish temporal coupling through selective attention, improving the precision of inter-regional information integration.

Speech understanding requires retrieving corresponding lexical information through phonetic features and combining them into phrases and sentences based on listeners' prior grammatical knowledge. Excluding prosodic and statistical cues, researchers found that different frequency cortical activities can simultaneously track temporal dynamics of abstract linguistic structures at word, phrase, and sentence levels. Synchronous neural entrainment to speech units at different temporal scales may 预示 a hierarchical embedding pattern, where representations of smaller speech units are embedded under representations of higher-level units, enabling timely integration of information across different hierarchical levels in speech.

When listeners understand speech content, context-based expectations can also influence neural entrainment to subsequent words' speech envelopes—the closer a word' s semantics is to context, the stronger the neural entrainment to that target word' s cortical EEG signals. This indicates that neural entrainment is also influenced by listeners' context-based predictions, maximizing predictability of future events and precisely timing resource allocation, thereby facilitating early encoding stages of upcoming words. This mechanism also explains why externally predictable rhythmic stimuli are more easily perceived than unpredictable non-rhythmic stimuli.

Top-down modulation of neural entrainment by listeners enables better representation of rhythmic information in complex auditory environments, promoting target speech understanding. Neural entrainment can serve as a "filter," reducing or eliminating neural responses to unattended speech streams in higher brain regions in noisy environments according to listeners' selective attention. It can also act as an "amplifier," enhancing representation and processing of corresponding speech components based on listeners' expectations. Finally, neural entrainment can function as a "connector," integrating information across different hierarchical components within speech or across brain regions according to listeners' prior knowledge. Thus, listeners' active regulation makes key information in speech understanding more likely to be at optimal excitability levels of neuronal ensemble activity, thereby receiving more processing resources. We therefore propose that neural entrainment may provide a "bridge" between

external and internal rhythms.

5. Discussion

5.1 External Rhythms Facilitate Speech Understanding

Inter-word pauses and pause locations in spoken language' s prosodic structure rhythm affect listeners' intelligibility and structural analysis of ambiguous contexts. Appropriate prosodic structure rhythm promotes correct speech understanding and restores semantically incomprehensible content. Different speech rates in context alter listeners' subsequent syllable discrimination and even word quantity perception. Additionally, speakers' synchronized motor behaviors during speech production transmit simultaneously with speech information to listeners' brains through visual channels. The synergy between these non-auditory motor rhythms and speech rhythms helps listeners better capture target speech content. Speech understanding benefits from these external rhythmic features, which not only aid comprehension and reduce processing costs but also modulate phoneme-, word-, and sentence-level speech processing.

When studying speech rhythm using speech material duration, the duration of acoustic units in spoken language changes perceived speech rate. Speech rate typically alters rhythm perception by changing the percentage of vocalic intervals (%V) and the standard deviation of consonantal intervals (deltaC). However, this phenomenon is not universal across all languages; for example, speech rate changes in French do not affect deltaC' s coefficient of variation. Therefore, whether speech rate changes directly influence speech rhythm perception remains controversial across languages, suggesting that research involving acoustic unit duration must carefully control speech rate manipulation.

5.2 Neural Entrainment–A Possible Mechanism Linking Internal and External Rhythms

Revelations of brain neural activity have led researchers to propose that internal rhythmic neural oscillations represent speech signals, enabling listeners to process key information in signals, confirmed in syllable perception, semantic processing, and syntactic comprehension. Numerous recent studies have found that brain neural oscillations may exhibit entrainment to external rhythmic stimuli. Since the phase of ongoing neural activity reflects rhythmic fluctuations in neuronal excitability, when entrainment occurs, aligned neural activity and external stimuli can stably adjust processing gain for input stimuli. We therefore propose that neural entrainment is a possible mechanism linking internal and external rhythms in speech understanding.

Neural entrainment widely exists in processes where external rhythms influence speech understanding, providing a pathway for how the brain represents hierarchical information in speech. Entrainment to speech stimuli does not occur in a specific frequency band alone; from gamma band responding to acoustic

features, to theta band tracking speech temporal envelopes, or lower delta band for characters, words, and sentences in Chinese, the brain produces corresponding neural oscillations for entrainment at different hierarchical levels. Neural entrainment also explains how prosodic structure rhythm or contextual rhythm influences current speech understanding through its self-sustaining characteristics—entrainment generated by previously input rhythmic stimuli can persist for some time after stimulus change, affecting processing of current speech stimuli. Entrainment to body speech facilitates timed inter-regional communication, ensuring precise integration of motor and speech information.

Top-down modulation of neural entrainment by listeners provides physiological explanations for the roles of selective attention, prior knowledge, and expectation in speech understanding. Selective attention enables neural ensembles at high excitability levels to more concentratedly represent target stimuli through neural entrainment, thereby improving target speech identification. Conversely, neural activity hinders representation establishment for sensory stimuli that cannot align, as they randomly amplify or attenuate information. Prior grammatical knowledge achieves precise inter-level integration by simultaneously entraining to different hierarchical units in speech. When listeners understand context, expectations for upcoming words can enhance entrainment strength during word processing, facilitating early phonetic encoding. We therefore propose that neural entrainment is not merely passive brain activity responding to external rhythmic stimuli but can create an appropriate processing environment for current speech understanding according to listeners' cognitive states. As a metric quantifying consistency between two rhythmic activities, it has become a method for describing the bidirectional relationship between external speech and the brain, allowing researchers to investigate how rhythms or cognitive processes affect speech understanding.

5.3 Existing Challenges

Long-standing controversy exists regarding whether brain responses to sensory stimuli relate to intrinsic, ongoing neural oscillations, and direct evidence is lacking for whether neural entrainment is generated by neural oscillations. Researchers need to rigorously determine whether observed entrainment phenomena result from coupling between external stimuli and intrinsic neural oscillations or from a series of stimulus-evoked potentials. In many cases, so-called entrainment may merely reflect a series of evoked neuronal responses caused by regular sound input rather than true neural oscillations.

With advances in non-invasive brain stimulation technology, researchers are no longer limited to passively recording brain activity but have begun using external interventions to investigate neural oscillations' influence on speech understanding. Transcranial alternating current stimulation (tACS), unlike transcranial magnetic stimulation (TMS), is a completely silent stimulation method that excludes interference from extraneous experimental sounds. When tACS is applied to listeners' temporal regions during speech processing, interfering

with theta-band neural oscillation activity disrupts neural entrainment and decreases speech intelligibility performance. Conversely, when speech stimulus envelopes serve as electrical stimulation modalities, scalp stimulation can improve listeners' speech understanding in noisy environments. tACS aligned in frequency and phase with rhythmic auditory stimuli facilitates perception of continuous auditory events in auditory cortex. Future research, whether controlling synchronization or desynchronization between neural oscillations and external stimuli, will help provide more direct evidence for neural oscillations' role in speech perception. This "experimental" manipulation of brain oscillations allows determination of whether brain oscillations causally drive brain function rather than being epiphenomenal, by examining functional outcomes.

6. Conclusion

Auditory speech understanding involves multi-scale participation of internal and external rhythms. We first revealed how external rhythms influence auditory speech understanding through three common types: prosodic structure rhythm, contextual rhythm, and speaker body language rhythm. Second, we described the role of listeners' internal neural oscillations and neural entrainment phenomena in speech understanding. Finally, based on neural entrainment' s modulation by listeners' top-down cognitive processes, we discussed the possibility that neural entrainment serves as a key mechanism linking internal and external rhythms.

References

Abbs, J. H., Gracco, V. L., & Cole, K. J. (1984). Control of Multimovement coordination: Sensorimotor mechansims in speech motor programming. Journal of Motor Behavior, 16(2), 195-231. https://doi.org/https://doi.org/10.1080/00222895.1984.10735318

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proceedings of the National Academy of Sciences of the United States of America, 98(23), 13367-13372. https://doi.org/10.1073/pnas.201400998

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. Trends in Cognitive Sciences, 16(7), 390-398. https://doi.org/10.1016/j.tics.2012.05.003

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-Term Temporal Tracking of Speech Rate Affects Spoken-Word Recognition. Psychological Science, 25(8), 1546-1553. https://doi.org/10.1177/0956797614533705

Baltus, A., & Herrman, C. S. (2016). The importance of individual frequencies of endogenous brain oscillations for auditory cognition - A short review. Brain Research, 1640, 243-250. https://doi.org/10.1016/j.brainres.2015.09.030

Bishop, G. H. (1933). Cyclic changes in excitability of the optic pathway of the rabbit. American Journal of Physiology, 103(1), 213-224. https://doi.org/https://doi.org/10.1152/ajplegacy.1932.103.1.213

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. Attention Perception & Psychophysics, 79(1), 333-343. https://doi.org/10.3758/s13414-016-

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. Language Cognition and Neuroscience, 33(8), 955-967. https://doi.org/10.1080/23273798.2018.1439179

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. Proceedings of the Royal Society B-Biological Sciences, 288(1943). https://doi.org/10.1098/rspb.2020.2419

Bosker, H. R., Peeters, D., & Holler, J. (2020). How visual cues to speech rate influence speech perception. Quarterly Journal of Experimental Psychology, 73(10), 1523-1536. https://doi.org/10.1177/1747021820914564

Bosker, H. R., Sjerps, M. J., & Reinisch, E. (2020). Temporal contrast effects in human speech perception are immune to selective attention. Scientific Reports, 10(1), Article 5607. https://doi.org/10.1038/s41598-020-62613-8

Bourguignon, M., De Tiege, X., Op de Beeck, M., Ligot, N., Paquier, P., Van Bogaert, P., . . . Jousmaki, V. (2013). The pace of prosodic phrasing couples the listener' s cortex to the reader' s voice. Human Brain Mapping, 34(2), 314-326. https://doi.org/10.1002/hbm.21442

Breska, A., & Deouell, L. Y. (2017). Neural mechanisms of rhythm-based temporal prediction: Delta phase-locking reflects temporal predictability but not rhythmic entrainment. Plos Biology, 15(2), Article e2001665. https://doi.org/10.1371/journal.pbio.2001665

Bridwell, D. A., Henderson, S., Sorge, M., Plis, S., & Calhoun, V. D. (2018). Relationships between alpha oscillations during speech preparation and the listener N400 ERP to the produced speech. Scientific Reports, 8, Article 12838. https://doi.org/10.1038/s41598-018-31038-9

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Current Biology, 28(24), 3976-+. https://doi.org/10.1016/j.cub.2018.10.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Current Biology, 28(5), 803-+. https://doi.org/10.1016/j.cub.2018.01.080

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. The Journal of Neuroscience, 39(38), 7564-7575. https://doi.org/10.1523/jneurosci.0584-19.2019

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica, 49(3-4), 155-180. https://doi.org/10.1159/000261913

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. Psychonomic Bulletin & Review, 18(6), 1189-1196. https://doi.org/10.3758/s13423-011-0167-9

Buzsaki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. Science, 304(5679), 1926-1929. https://doi.org/10.1126/science.1099745

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. Trends in Cognitive Sciences, 18(6), 300-309. https://doi.org/10.1016/j.tics.2014.02.005

Cason, N., & Schoen, D. (2012). Rhythmic priming enhances the phonological processing of speech. Neuropsychologia, 50(11), 2652-2658. https://doi.org/10.1016/j.neuropsychologia.2012.07.018

Christiansen, M. H., & Chater, N. (2015). The now-or-Never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39. https://doi.org/10.1017/s0140525x1500031x

Cho, T., Whalen, D. H., & Docherty, G. (2019). Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. Journal of Phonetics, 72, 52-65. https://doi.org/10.1016/j.wocn.2018.11.002

Dauer, R. M. (1983). Stress-Timing and Syllable-Ttimg Reanalyzed. Journal of Phonetics, 11(1), 51-62. https://doi.org/10.1016/s0095-4470(19)30776-4

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), Language and language processing: Proceedings of the 38th linguistic colloquium (pp. 231–241).

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. International Congress of Phonetic Sciences, Barcelona/Spain. https://doi.org/10.5167/uzh-111779

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigne, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage, 196, 237-247. https://doi.org/10.1016/j.neuroimage.2019.04.037

Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. Journal of Memory and Language, 63(3), 274-294. https://doi.org/10.1016/j.jml.2010.06.003

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. Journal of Memory and Language, 59(3), 294-311. https://doi.org/10.1016/j.jml.2008.06.006

Dilley, L. C., & Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. Psychological Science, 21(11), 1664-1670. https://doi.org/10.1177/0956797610384743

Ding, N., & He, H. (2016). Rhythm of Silence. Trends in Cognitive Sciences, 20(2), 82-84. https://doi.org/10.1016/j.tics.2015.12.006

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00481

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience, 19(1), 158-+. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. Neuroscience and Biobehavioral Reviews, 81, 181-187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., & Zhang, J. (2018). Attention is required for knowledge-based sequential grouping: Insights from the integration of syllables into words. The Journal of Neuroscience, 38(5), 1178-1188. https://doi.org/10.1523/jneurosci.2606-17.2017

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophysiology, 107(1), 78-89. https://doi.org/10.1152/jn.00297.2011

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage, 85, 761-768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Doelling, K. B., & Assaneo, M. F. (2021). Neural oscillations are a start toward understanding brain activity rather than the end. PLOS Biology, 19(5), e3001234. https://doi.org/10.1371/journal.pbio.3001234

Farbood, M. M., Marcus, G., & Poeppel, D. (2013). Temporal Dynamics and the Identification of Musical Key. Journal of Experimental Psychology-Human Perception and Performance, 39(4), 911-918. https://doi.org/10.1037/a0031087

Feher, K. D., Nakataki, M., & Morishima, Y. (2017). Phase-Dependent Modulation of Signal Transmission in Cortical Networks through tACS-Induced Neural Oscillations. Frontiers in Human Neuroscience, 11, Article 471. https://doi.org/10.3389/fnhum.2017.00471

Fiedler, L., Woestmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. Neuroimage, 186, 33-42. https://doi.org/10.1016/j.neuroimage.2018.10.057

Fuglsang, S. A., Dau, T., & Hjortkjaer, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage, 156, 435-444. https://doi.org/10.1016/j.neuroimage.2017.04.026

Fujii, S., & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: the SEP hypothesis. Frontiers in Human Neuroscience, 8, Article 777. https://doi.org/10.3389/fnhum.2014.00777

Ghazanfar, A. A., & Takahashi, D. Y. (2014). The evolution of speech: vision, rhythm, cooperation. Trends in Cognitive Sciences, 18(10), 543-553. https://doi.org/10.1016/j.tics.2014.06.004

Ghitza, O., & Greenberg, S. (2009). On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence. Phonetica, 66(1-2), 113-126. https://doi.org/10.1159/000208934

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience, 15(4), 511-517. https://doi.org/10.1038/nn.3063

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., . . . Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party" . Neuron, 77(5), 980-991. https://doi.org/10.1016/j.neuron.2012.12.037

Haegens, S., & Golumbic, E. Z. (2018). Rhythmic facilitation of sensory processing: A critical review. Neuroscience and Biobehavioral Reviews, 86, 150-165. https://doi.org/10.1016/j.neubiorev.2017.12.002

Helfrich, R. F., Breska, A., & Knight, R. T. (2019). Neural entrainment and network resonance in support of top-down guided attention. Current Opinion in Psychology, 29, 82-89. https://doi.org/10.1016/j.copsyc.2018.12.016

Henry, M. J., Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. Proceedings of the National Academy of Sciences of the United States of America, 111(41), 14935-14940. https://doi.org/10.1073/pnas.1408741111

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. Trends in Cognitive Sciences, 23(8), 639-652. https://doi.org/10.1016/j.tics.2019.05.006

Iani, F., & Bucciarelli, M. (2017). Mechanisms underlying the beneficial effect of a speaker' s gestures on the listener. Journal of Memory and Language, 96, 110-121. https://doi.org/10.1016/j.jml.2017.05.004

Jadoul, Y., Ravignani, A., Thompson, B., Filippi, P., & de Boer, B. (2016). Seeking Temporal Predictability in Speech: Comparing Statistical Approaches on 18 World Languages. Frontiers in Human Neuroscience, 10. https://doi.org/10.3389/fnhum.2016.00586

Jensen, O., Bonnefond, M., & VanRullen, R. (2012). An oscillatory mechanism for prioritizing salient unattended stimuli. Trends in Cognitive Sciences, 16(4), 200-206. https://doi.org/10.1016/j.tics.2012.03.002

Kayser, C. (2019). Evidence for the Rhythmic Perceptual Sampling of Auditory Scenes. Frontiers in Human Neuroscience, 13, https://doi.org/10.3389/fnhum.2019.00249

Kayser, C., Wilson, C., Safaai, H., Sakata, S., & Panzeri, S. (2015). Rhythmic auditory cortex activity at multiple timescales shapes stimulus-response gain and background firing. Journal of Neuroscience, 35(20), 7750-7762. https://doi.org/10.1523/jneurosci.0268-15.2015

Keshavarzi, M., & Reichenbach, T. (2020). Transcranial Alternating Current Stimulation With the Theta-Band Portion of the Temporally-Aligned Speech Envelope Improves Speech-in-Noise Comprehension. Frontiers in Human Neuroscience, 14, https://doi.org/10.3389/fnhum.2020.00187

Knudsen, E. I. (2018). Neural Circuits That Mediate Selective Attention: A Comparative Perspective. Trends in Neurosciences, 41(11), 789-805. https://doi.org/10.1016/j.tins.2018.06.006

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural Entrainment Determines the Words We Hear. Current Biology, 28(18), 2867-2875. https://doi.org/10.1016/j.cub.2018.07.023

Kösem, A., & van Wassenhove, V. (2016). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Language Cognition and Neuroscience, 32(5), 536-544. https://doi.org/10.1080/23273798.2016.1238495

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The Evolution of Rhythm Processing. Trends in Cognitive Sciences, 22(10), 896-910. https://doi.org/10.1016/j.tics.2018.08.002

Kotz, S. A., & Schmidt-Kassow, M. (2015). Basal ganglia contribution to rule expectancy and temporal predictability in speech. Cortex, 68, 48-60. https://doi.org/10.1016/j.cortex.2015.02.021

Kotz, S. A., & Schwartze, M. (2010). Cortical speech processing unplugged: a timely subcortico-cortical framework. Trends in Cognitive Sciences, 14(9), 392-399. https://doi.org/10.1016/j.tics.2010.06.005

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31(1), 32-59. https://doi.org/10.1080/23273798.2015.1102299

Ladefoged, P. (1975). A Course in Phonetics. Harcourt Brace Jovanovich College.

Lakatos, P., Gross, J., & Thut, G. (2019). A New Unifying Account of the Roles of Neuronal Entrainment. Current Biology, 29(18), R890-R905. https://doi.org/10.1016/j.cub.2019.07.075

Lakatos, P., O' Connell, M. N., Barczak, A., Mills, A., Javitt, D. C., & Schroeder, C. E. (2009). The Leading Sense: Supramodal Control of Neurophysiological Context by Attention. Neuron, 64(3), 419-430. https://doi.org/10.1016/j.neuron.2009.10.014

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. Journal of Neurophysiology, 94(3), 1904-1911. https://doi.org/10.1152/jn.00263.2005

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. Journal of Experimental Psychology-Human Perception and Performance, 21(3), 451-468. https://doi.org/10.1037/0096-1523.21.3.451

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive Windows using a narrated story. Journal of Neuroscience, 31(8), 2906-2915. https://doi.org/10.1523/jneurosci.3684-10.2011

Li, W., & Yang, Y. (2009). Perception of prosodic hierarchical boundaries in Mandarin Chinese Sentences. Neuroscience, 158(4), 1416-1425. https://doi.org/10.1016/j.neuroscience.2008.10.065

Li, W., & Yang, Y. (2010). Perception of Chinese poem and its electrophysiological effects. Neuroscience, 168(3), 757-768. https://doi.org/10.1016/j.neuroscience.2010.03.069

Li, W., Zhang, H., Zheng, Z., & Li, X. (2019). Prosodic phrase priming during listening to Chinese ambiguous phrases in different experimental tasks. Journal of Neurolinguistics, 51, 135-150. https://doi.org/10.1016/j.jneuroling.2019.02.003

Li, X., & Ren, G. (2012). How and when accentuation influences temporally selective attention and subsequent semantic processing during on-line spoken language comprehension: An ERP study. Neuropsychologia, 50(8), 1882-1894. https://doi.org/10.1016/j.neuropsychologia.2012.04.013

Li, X., Shao, X., Xia, J., & Xu, X. (2019). The cognitive and neural oscillatory mechanisms underlying the facilitating effect of rhythm regularity on speech comprehension. Journal of Neurolinguistics, 49, 155-167. https://doi.org/10.1016/j.jneuroling.2018.05.004

Ling, L. E., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. Language and Speech, 43, 377-401. https://doi.org/10.1177/00238309000430040301

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron, 54(6), 1001-1010. https://doi.org/10.1016/j.neuron.2007.06.004

Luo, Y., Duan, Y., & Zhou, X. (2015). Processing rhythmic pattern during Chinese sentence reading: An eye movement study. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.01881

Luo, Y., & Zhou, X. (2010). ERP evidence for the online processing of rhythmic pattern during Chinese sentence reading. NeuroImage, 49(3), 2836-2849. https://doi.org/10.1016/j.neuroimage.2009.10.008

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., & Golumbic, E. Z. (2017). Sleep Disrupts High-Level Speech Parsing Despite Significant Basic Auditory Processing. Journal of Neuroscience, 37(32), 7772-7781. https://doi.org/10.1523/jneurosci.0168-17.2017

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the Tracking of Habitual Rate Influences Speech Perception. Journal of Experimental Psychology-Learning Memory and Cognition, 45(1), 128-138. https://doi.org/10.1037/xlm0000579

Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M., & Lleras, A. (2010). Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. Cognition, 115(1), 186-191. https://doi.org/10.1016/j.cognition.2009.11.010

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233-U118. https://doi.org/10.1038/nature11020

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. Proceedings of the National Academy of Sciences of the United States of America, 114(42), E8913-E8921. https://doi.org/10.1073/pnas.1705373114

Morillon, B., Schroeder, C. E., & Wyart, V. (2014). Motor contributions to the temporal precision of auditory attention. Nature Communications, 5, Article 5255. https://doi.org/10.1038/ncomms6255

Morris, D. J., & Klerke, S. (2016). Machine classification of P1-N1-P2 responses elicited with a gated syllable. The Journal of the Acoustical Society of America, 140(4), 3155-3155. https://doi.org/https://doi.org/10.1121/1.4969899

Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D., & Tessendorf, S. (2013). Body-language-communication: . In An international handbook on multimodality in human interaction (Vol. 1, pp. 131-232). De Gruyter Mouton.

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), The Handbook of the phonetic sciences (Vol. 5, pp. 640-673). Blackwell Publishers.

O' Brien, G. E., Gijbels, L., & Yeatman, J. D. (2020). Context effects on phoneme categorization in children with dyslexia. Journal of the Acoustical Society of America, 148(4), 2209-2222. https://doi.org/10.1121/10.0002181

Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain. Trends in Cognitive Sciences, 23(11), 913-926. https://doi.org/10.1016/j.tics.2019.08.004

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife, 5. https://doi.org/10.7554/elife.14521

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. Current Biology, 25(12), 1649-1653. https://doi.org/10.1016/j.cub.2015.04.049

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology, 3, https://doi.org/10.3389/fpsyg.2012.00320

Pike, K. L. (1945). The Intonation of American English.

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. Attention Perception & Psychophysics, 78(1), 334-345. https://doi.org/10.3758/s13414-015-0981-7

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. Nature Reviews Neuroscience, 21(6), 322-334. https://doi.org/10.1038/s41583-020-0304-4

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1493), 1071-1086. https://doi.org/10.1098/rstb.2007.2160

Proctor, M., Walker, R., Smith, C., Szalay, T., Goldstein, L., & Narayanan, S. (2019). Articulatory characterization of English liquid-final rimes. Journal of Phonetics, 77, Article 100921. https://doi.org/10.1016/j.wocn.2019.100921

Raco, V., Bauer, R., Tharsan, S., & Gharabaghi, A. (2016). Combining TMS and tACS for Closed-Loop Phase-Dependent Modulation of Corticospinal Excitability: A Feasibility Study. Frontiers in Cellular Neuroscience, 10, https://doi.org/10.3389/fncel.2016.00143

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives Proc Speech Prosody, Aix-en-Provence.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265-292. https://doi.org/10.1016/s0010-0277(99)00058-x

Reichle, M. E. (2010). Two views of brain function. Trends in Cognitive Sciences, 14(4), 180-190. https://doi.org/10.1016/j.tics.2010.01.008

Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. Attention Perception & Psychophysics, 78(4), 1203-1217. https://doi.org/10.3758/s13414-016-1067-x

Riecke, L., Formisano, E., Sorger, B., Baskent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. Current Biology, 28(2), 161-169. https://doi.org/10.1016/j.cub.2017.11.033

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive Sensing of Periodic and Aperiodic Auditory Patterns. Trends in Cognitive Sciences, 22(10), 870-882. https://doi.org/10.1016/j.tics.2018.08.003

Roach, P. (1982). On the distinction between ' stress-timed' and ' syllable-timed' languages. Linguistic controversies, 73, 79.

Rohenkohl, G., Cravo, A. M., Wyart, V., & Nobre, A. C. (2012). Temporal Expectation Improves the Quality of Sensory Information. Journal of Neuroscience, 32(24), 8424-8428. https://doi.org/10.1523/jneurosci.0804-12.2012

Schmidt-Kassow, M., Roncaglia-Denissen, M. P., & Kotz, S. A. (2013). Speech Rhythm Facilitates Syntactic Ambiguity Resolution: ERP Evidence. Figshare. https://doi.org/https://doi.org/10.1371/journal.pone.0056000

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences, 32(1), 9-18. https://doi.org/10.1016/j.tins.2008.09.012

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends in Cognitive Sciences, 12(3), 106-113. https://doi.org/10.1016/j.tics.2008.01.002

Sheng, J., Zheng, L., Lyu, B., Cen, Z., Qin, L., Tan, L. H., . . . Gao, J.-H. (2019). The Cortical Maps of Hierarchical Linguistic Structures during Speech Perception. Cerebral Cortex, 29(8), 3232-3240. https://doi.org/10.1093/cercor/bhy191

Steinmetzger, K., & Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. Neuropsychologia, 95, 173-181. https://doi.org/10.1016/j.neuropsychologia.2016.12.003

Stilp, C. (2020). Acoustic context effects in speech perception. Wiley Interdisciplinary Reviews-Cognitive Science, 11(1), Article e1517. https://doi.org/10.1002/wcs.1517

Tass, P., Rosenblum, M. G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., . . . Freund, H. J. (1998). Detection of n : m phase locking from noisy data: Application to magnetoencephalography. Physical Review Letters, 81(15), 3291-3294. https://doi.org/10.1103/PhysRevLett.81.3291

Thut, G., Schyns, P. G., & Gross, J. (2011). Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. Frontiers in Psychology, 2, Article 170. https://doi.org/10.3389/fpsyg.2011.00170

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapic, and Goswami and Leong. Labora-

tory Phonology, 4(1), 93-118. https://doi.org/10.1515/lp-2013-0005

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. Jaro-Journal of the Association for Research in Otolaryngology, 19(2), 181-191. https://doi.org/10.1007/s10162-018-0654-z

Vosskuhl, J., Strueber, D., & Herrmann, C. S. (2018). Non-invasive Brain Stimulation: A Paradigm Shift in Understanding Brain Oscillations. Frontiers in Human Neuroscience, 12, Article 211. https://doi.org/10.3389/fnhum.2018.00211

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. Perception & Psychophysics, 67(6), 939-950. https://doi.org/10.3758/bf03193621

Wang, M., Kong, L., Zhang, C., Wu, X., & Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions. Journal of the Acoustical Society of America, 143(4), EL255-EL259. https://doi.org/10.1121/1.5030518

White, L. (2014). Communicative function and prosodic form in speech timing. Speech Communication, 63-64, 38-54. https://doi.org/10.1016/j.specom.2014.04.003

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. Journal of Memory and Language, 66(4), 665-679. https://doi.org/10.1016/j.jml.2011.12.010

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. Neuroimage, 172, 766-774. https://doi.org/10.1016/j.neuroimage.2018.01.038

Wu, C., Cao, S., Wu, X., & Li, L. (2013). Temporally pre-presented lipreading cues release speech from informational masking. Journal of the Acoustical Society of America, 133(4), EL281-EL285. https://doi.org/10.1121/1.4794933

Zhang, W., & Ding, N. (2017). Time-domain analysis of neural tracking of hierarchical linguistic structures. Neuroimage, 146, 333-340. https://doi.org/10.1016/j.neuroimage.2016.11.016

Zion-Golumbic, E., & Schroeder, C. E. (2012). Attention modulates 'speech-tracking' at a cocktail party. Trends in Cognitive Sciences, 16(7), 363-364. https://doi.org/10.1016/j.tics.2012.05.004

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. Current Biology, 28(3), 401-408. https://doi.org/10.1016/j.cub.2017.11.071

方岚, 郑苑仪, 金晗, 李晓庆, 杨玉芳, & 王瑞明. (2021). 口语句子的韵律边界: 窥探言语理解的秘窗. 心理科学进展, 29(3), 425-437. https://doi.org/https://dx.doi.org/10.3724/SP.J.1042.2021.00425

杨玉芳. (2021). 语言理解——认知过程和神经基础. 科学出版社.

殷融. (2020). “动手不动口”：手部动作与语言进化的关系. 心理科学进展, 28(7), 1141-1155. https://doi.org/10.3724/SP.J.1042.2020.01141

于泽, 韩玉昌, & 任桂琴. (2010). 韵律在语言加工中的作用及其神经机制. 心理科学进展, 18(3), 420-425.

Note: Figure translations are in progress. See original paper for figures.