# Deciphering the Rhythmic Symphony of Speech: A Neural Framework for Robust and Time-Invariant Speech Comprehension

Olesia Dogonasheva[1,2*], Denis Zakharov[2], Anne-Lise Giraud[3], Boris Gutkin[1]

[1]Group of Neural Theory, École Normale Supérieure PSL*, Paris, France.
[2]Centre for Cognition and Decision Making, HSE University, Moscow, Russia.
[3]Institut de l'Audition, The Institut Pasteur, Université de Paris Cité, Paris, France.

*Corresponding author(s). E-mail(s): odogonasheva@gmail.com;

**Abstract**

Unraveling the mysteries of how humans effortlessly grasp speech amidst diverse environmental challenges has long intrigued researchers in systems and cognitive neuroscience. This study delves into the neural intricacies underpinning robust speech comprehension, giving a computational mechanistic proof for the hypothesis proposing a pivotal role for rhythmic, predictive top-down contextualization facilitated by the delta rhythm in achieving time-invariant speech processing. We propose a Brain-Rhythm-Based Inference (BRyBI) model that integrates three key rhythmic processes – theta-gamma interactions for parsing phoneme sequences, dynamic delta rhythm for inferred prosodic-phrase context, and resilient speech representations. Demonstrating mechanistic proof-of-principle, BRyBI replicates human behavioral experiments, showcasing its ability to handle pitch variations, time-warped speech, interruptions, and silences in non-comprehensible contexts. Intriguingly, the model aligns with human experiments, revealing optimal silence time scales in the theta- and delta-frequency ranges. Comparative analysis with deep neural network language models highlights distinctive performance patterns, emphasizing the unique capabilities of our rhythmic framework. In essence, our study sheds light on the neural underpinnings of speech processing, emphasizing the role of rhythmic brain mechanisms in structured temporal signal processing – an insight that challenges prevailing

artificial intelligence paradigms and hints at potential advancements in compact and robust computing architectures.

**Keywords:** rhythms, predictive coding, speech recognition, inference model, invariant speech processing, auditory cortex

# 1 Introduction

Speech processing, with its inherent complexities and multidimensional nature, continues to be a focal point of cognitive neuroscience. Humans possess an extraordinary ability to comprehend speech across a wide spectrum of voices, ranging from young children to elderly individuals, from speakers of different languages to regional dialects, and even across diverse socio-cultural backgrounds. Moreover, speech comprehension remains robust despite variations in speech rates, encompassing both rapid and leisurely speech patterns.

However, the invariance of speech processing extends beyond robustness to different voices spoken in noisy conditions. In multiple studies, speech comprehension was found to be largely impervious to manipulations of speech structure, including interruptions and segmentations. In experiments with interrupted speech [1], silent intervals masked the speech at different time frequencies, i.e., the speech signal was interrupted by silences. As a result, some elements of the speech were simply missing. The results of the experiment showed that when the frequency of the interruptions was greater than 1 Hz (1000 ms of signal, 1000 ms of silence), speech recognition recovered to nearly control levels.

In another set of studies, silent intervals of different durations (up to 500 ms) were inserted into the speech. Here, the segmented signal contained all the parts of the original speech, and no information was deleted [2]. In these experiments, subjects' performance showed characteristic U-shaped curves, with the worst performance when the silence durations were 100 ms, whatever the silence-to-speech rate. In another manipulation, speech was compressed by different factors. Subjects in these tasks showed robust success in recognising speech as long as the compression factor was less than 2. Speech comprehension dropped catastrophically when compression factors were above 2 [3, 4]. Intriguingly, when this temporally squeezed and incomprehensible speech was split into chunks interspersed with silences, recognition recovered [5]. Here, the performance errors showed a characteristic U-shape with the fewest errors when the overall natural duration of speech was restored by the silent insertions. These results underline the importance of aligning the temporal scales of speech with endogenous scales set by multiple brain rhythms in reconstructing meaning from the acoustic speech flow. Understanding the mechanisms that enable humans to navigate through this large parameter space presents an intriguing challenge and arguably, a litmus test for the potential neural mechanisms underlying speech recognition processes. Notably, how could we explain why speech comprehension is recovered by adding silences that do not carry any information?

We propose that a compelling explanation could be constructed by leveraging a prominent hypothesis asserting that the hierarchical rhythmic structure inherent in meaningful speech significantly contributes to robust and temporally invariant comprehension [6–9]. In fact, the hierarchical configuration of natural language, which in turn constrains the structure of speech, appears to be pivotal for the efficiency of human speech processing [10–14]. The hierarchical organization of speech spans all linguistic levels, from the phonetic structure of words to the highest tiers of communication [15]. Phonemes coalesce into syllables and words, while the syntactic hierarchy orchestrates the assembly of words into phrases, further evolving into sentences. This line of reasoning prompts a fundamental inquiry: What neural mechanisms underpin the harnessing of the hierarchical nature of speech for achieving effective processing, encompassing tasks such as parsing and comprehension?

Brain rhythms emerge as a compelling candidate for the neural mechanisms, capitalizing on the intrinsic hierarchical rhythmic organization of speech [16–19]. Substantial empirical evidence supports the notion that rhythmic brain activity maintains a hierarchical structure during the processing of speech, and this hierarchy aligns with the inherent structure of speech itself [20–22]. It is thus plausible to posit that the rhythmic structure of speech interacts with the scaffold of endogenous brain rhythms, thereby establishing temporal processing windows. These windows, in turn, facilitate the real-time processing and comprehension of incoming auditory signals [19, 23].

A natural mechanism facilitating temporal windowing involves the entrainment or synchronization of neural activity with a rhythmic stimulus, such as speech. In the primary auditory cortex, for instance, the theta rhythm is acknowledged to be entrained by the speech envelope, thereby encoding syllabic information [23, 24]. Concurrently, oscillations in the gamma range embedded within a theta-cycle have been demonstrated to encode phonemes [25, 26], giving rise to a theta-gamma code for syllables [17, 27]. Previous investigations have proposed that the theta-gamma code orchestrates a bottom-up information flow, commencing from sounds captured by the cochlea and converging in the primary auditory cortex [5, 28, 29]. This rhythmic windowing, characterized by theta-gamma dynamics, may confer robustness to speech parsing in noisy and compressed speech scenarios [5], with comprehension recovery contingent on the reinstatement of the natural syllabic rate based on feedforward gamma-coding [29, 30].

However, this conceptual model falls short in elucidating how such recovery aligns with human performance in experiments involving perturbed speech comprehension under interruptions and segmentation [1, 2]. We propose that a comprehensive explanation needs to take into account the influence of top-down factors like semantics and context [31–33].

Hence, as a key conceptual proposition in this paper, we suggest that a top-down predictive information flow modulated by rhythm can mitigate? the deterioration of speech signals and improve processing reliability in acoustically challenging environments [34–39]. More specifically, we hypothesize that the information from multiple syllables is predictively combined into a semantic contextual representation (e.g., a word or a prosodic phrase) via a process indelibly intertwined with the delta rhythm

[40]. Nevertheless, the computational mechanisms governing the formation of such predictive representations and how this process distinctly contributes to speech processing in the brain remain open questions.

The delta rhythm, being the slowest rhythm observed in the auditory cortex during speech processing, is experimentally observed in various aspects. Studies suggest that the functions of the delta rhythm include tracking of prosody [41–45], chunking of words and phrases [46, 47], error resolution [48, 49], multiscale integration [40, 50] and top-down modulation of speech processing [35, 51, 52]. Existing models primarily employ the delta rhythm as a mechanism for chunking words and phrases [53, 54], overlooking testing its hypothesized role in top-down contextual influence.

To illustrate the integration of rhythm-based bottom-up signals with top-down contextual influences, forming resilient and consistent speech representations and processing, we propose the Brain-Rhythm-Based Inference model (BRyBI). BRyBI mechanistically incorporates diverse brain-rhythm data and accommodates time-invariant speech recognition. In this model, hierarchically organized interacting rhythms actively sustain the flow of both top-down and bottom-up information during the inference process: theta-gamma interactions delineate and parse the phoneme/syllable sequences, while the delta rhythm dynamically generates the inferred word/prosodic-phrase context. We demonstrate how these processes facilitate speech recognition even in complex conditions. Additionally, we elucidate the mechanisms underlying the remarkable recovery of comprehension of perturbed speech when specific time-scales of the spoken rhythm are re-established. Our contention is that rhythmic predictive top-down contextualization plays a pivotal role in explaining time-invariant speech processing. Furthermore, our model predicts the restoration of comprehension in compressed speech through the re-chunking of words and phrases, emphasizing the critical dependence on top-down delta-dependent processes.

# 2 Results

## 2.1 Conceptual structure of the rhythm-based bayesian inference computation for speech processing

Our proposed model is fundamentally rooted in the predictive coding framework, wherein prior information or beliefs are encoded within an internal model of the environment, often referred to as a generative model (GM) [55, 56]. This internal model actively influences perception [57]. The GM generates predictions of sensory signals, and these predictions, or beliefs, are subsequently compared with the actual incoming peripheral signals. The resultant comparison yields prediction errors that traverse the model hierarchy to update the internal states within the GM. Multiple studies have demonstrated that the predictive coding framework provides a plausible paradigm for audio perception [58]. Firstly, it facilitates the establishment of a hierarchical structure that emphasizes linguistic organization and the hierarchy of speech processing [14, 59]. Secondly, states in a predictive model are dynamic, reflecting the nature of brain processes. Thirdly, predictive coding integrates both top-down predictions and bottom-up mismatch errors. Lastly, it enables real-time speech parsing. Recent advances in predictive coding models have demonstrated a balanced implementation

of linguistic aspects and the mechanistic plausibility of biophysical algorithms for speech processing [29, 60–62]. Consequently, we have implemented the BRyBI model as a predictive coding model, wherein bottom-up and top-down rhythm-based processes are structured along a theta-based code for syllable parsing (bottom-up) and a delta-based top-down predictive code for phrase parsing and comprehension.

We conceive the rhythm-governed Bayesian inference in BRyBI to be architected as a two-level GM model. The bottom and top levels of the BRyBI parallel speech processing are in the primary auditory cortex (pAC) and the associative auditory cortex (aAC), respectively (Fig. 1). At the top level, the delta rhythm provides the temporal scaffold for semantic context formation, while coupled theta and gamma rhythms at the bottom level encode the acoustic signal of speech, depending on the context. The context represents the prosodic phrases and sets predictions for the sequence of the constituent syllables and phonemes.

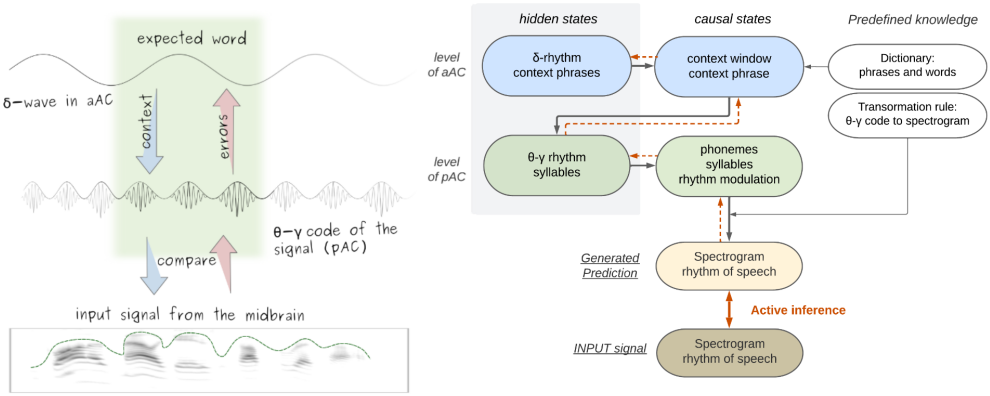


**Fig. 1** The BRyBI model incorporates predictive Bayesian inference for rhythm-based dynamical speech formation. (A) A conceptual illustration of the general BRyBI structure. The hierarchy comprises the level of the primary auditory cortex (pAC) and that of associative auditory cortices (aAC). At the aAC level, the delta rhythm governs semantic context formation as an expected word or prosodic phrase and passes it to pAC level, where coupled theta-gamma rhythms encode the acoustic signal conditioned by the context. The pAC level combines information from both top-down and bottom-up flows, creating a theta-code of speech. The theta-gamma code is transformed into a spectrogram that is compared with the input signal from the midbrain. The prediction error is passed bottom-up and used to infer the next prediction. Blue arrows represent the top-down flow in a generative model, and red arrows represent the bottom-up error flow. (B) Implementation scheme of the BRyBI GM. Hidden states are represented as dynamical variables (gray background), and causal states are their nonlinear transformation. The top level implements a context phrase and delta rhythm, which, after nonlinear transformation, provide a window for prosodic context formation. Information about context is then passed to pAC level and used for theta-gamma rhythm and syllable dynamics. At this level, phonemes, syllables, and modulation signals (rhythm of speech, analogue of speech envelope) are used to generate a spectrogram, which, together with the speech rhythm, is compared with the input signal. The generative model supports top-down information flow (black arrows), while the inference provides bottom-up error passing (red dashed arrows) through the hierarchy.

The delta rhythm provides the temporal scaffold for contextual information in the BRyBI model. In general, the role of the delta rhythm in speech processing is not well established. Current research points to the delta rhythm being entrained by the speech envelope [41–45]. Entrainment is considered a mechanism for word and phrase parsing [46, 47]. However, the delta rhythm can also endogenously arise from a theta-syllable string, thus being involved in a "chunking" process that creates word-level chunks from syllables units [18]. The delta rhythm is observed at a higher hierarchical level and can correlate with error resolution [48, 49], top-down information passing [35, 37, 51, 52] and multiscale integration [40, 50, 63]. Based on these proposals, we posit that the context of words and prosodic phrases is influenced by top-down mechanisms regulated by the delta rhythm [43, 53]. Implicitly, the BRyBI model implements delta rhythm for temporal segmentation and context formation.

We incorporate a theta-gamma code for syllables in our BRyBI model in accordance with numerous experiments [23, 24, 64–67], where the theta rhythm was shown to be entrained by the speech envelope and thus synchronized with syllables, and the gamma rhythm that is coupled with theta rhythm encodes phonemes. In the BRyBI model, the theta rhythm is also entrained by a rhythm of syllables and performs temporal segmentation of the continuous signal into the syllables. The coupled gamma rhythm is involved in phoneme coding (see Methods: Bottom Level). A similar realization of these mechanisms, where the coupling of theta and gamma rhythms improved speech processing, was proposed [29].

The syllable formation dynamics are rhythmically controlled by theta rhythm (Fig. 2D) in an interactive activation process [68, 69]. Decoded from the spectrogram, phonemes activate the corresponding elements in the network that represent syllables at the bottom level. At the same time, the top level of the processing hierarchy creates a pattern of possible and probable syllable sequence transitions for the current word or phrase. Let us consider an illustrative example, recognizing the sentence "This was easy for us". This sentence can be seen to consist of two phrases: "This was easy" and "for us" (Fig. 2A, Fig. 2D). The first set of phonemes activates the syllable "ðɪs". For the next two syllables, the system has zero uncertainty (100% of confidence), and the context dictates the activation of the syllables "wəz" and then "zi" even if the signal of phonemes is distorted or isn't received. For the next syllable, uncertainty is 50% (in this example, there are two ways: to activate a syllable "zi" again or to finish the phrase with a pause denoted by "#"), so here the system reconstructs the spectrogram more carefully. The theta rhythm is synchronized with a speech envelope more precisely for the bigger uncertainty (Fig. 2C). And finally, the system receives the next phoneme, "z", and can follow a certain context of the phrase "this was easy" with 100% confidence. This example of reconstruction is simplified as much as possible in order to demonstrate the mechanism behind it.

To drive the model, we use both the spectrogram and syllable/prosodic envelopes (high-pass and low-pass envelopes, respectively) as inputs to the bottom level. The original dataset is a preprocessed TIMIT dataset [70]. Using the matlab code [29], we extracted the 6-channel spectra from the sentences ass described in Section Methods: Dataset.

Once the model constructs a candidate speech signal, the DEM algorithm is used to infer and optimize the beliefs in the generative model [57]. Beliefs here are the states of all variables in GM, including candidates for current context (phrases and words), syllables, phonemes, and the phase of delta and theta rhythms (see details in Methods). During inference, the trajectories in the GM are reconstructed, and inferred phrases and syllables are compared to ground truth phrases and syllables (Fig. 2A).

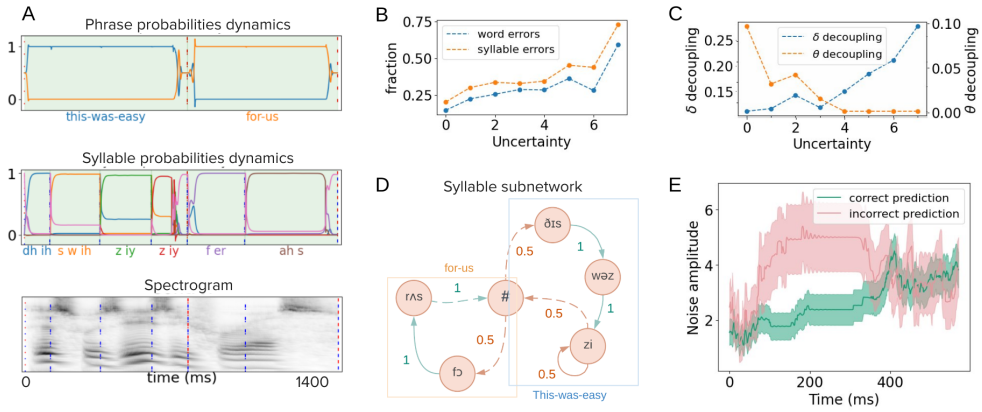


**Fig. 2** General performance of the BRyBI model for natural speech (A) An example of speech recognition by the model for the test sentence: "this was easy for us". This sentence consists of two prosodic phrases: "this was easy" and "for us". The top panel represents dynamics for phrase probabilities. During online speech recognition, BRyBI accumulates evidence for the current phrase and infers probability for each phrase at the top level. Similar accumulation happens at the bottom level for syllable probabilities (middle panel). The green background represents correctly recognized phrases and syllables. The red background indicates errors. Line colors correspond to phrases and syllables (signed on x-axes). The bottom panel shows the corresponding spectrogram. (B) The fraction of errors in words and syllables depends on uncertainty (represented as numbers of increasing intervals; see Supplementary for details). As uncertainty grows, speech recognition errors increase, though growth is bounded at an uncertainty unit value of 6 due to the theta rhythm locking mechanism. (C) Rhythm decoupling depends on uncertainty. With increased uncertainty, stronger theta coupling with the speech envelope enables reliable phoneme decoding, facilitating speech recognition in unpredictable/noisy conditions. At low certainty levels, the context aids predictions (delta-speech coupling relatively pronounced), but theta rhythm locking is not needed. (D) An example of a syllable network that is implemented at the bottom level of BRyBI. Each node represents a syllable, and connections represent possible transitions between syllables. Weights on connections represent confidence in the transition. The sign "#" is a pause that separates words and phrases. (E) Surprise as ERP in BRyBI. In cases of incorrect choice of phrase with high confidence (low uncertainty), a prediction error is passed bottom up and causes a change in the chosen phrase. The deviation from the chosen trajectory of dynamic variables occurs at aAC level due to noise addition. The noise amplitude correlates with the deviation and, thus, correlates with an error in the chosen semantic context.

## 2.2 Rhythm-regulated generative Bayesian inference model (BRyBI) recovers speech despite temporal and content perturbations

Speech recognition by the BRyBI model demonstrates good accuracy (15% word errors for 100 sentences) for natural speech input (see Fig. 2A for an example). The top-down process plays a crucial role in shaping the selection of subsequent syllables and phonemes in the model. When context is poorly established, predicting the next syllable becomes challenging, requiring increased sensitivity in theta-syllable synchronization [49, 71]. Such interplay between rhythms is possible through the predictive coding framework. Figures 2B and 2C demonstrate the model's performance in speech recognition and rhythm entrainment, depending on the uncertainty of the next phoneme. For higher uncertainty, the theta rhythm follows a syllabic rhythm precisely, enabling robust speech perception. As the uncertainty of the future phoneme increases, the speech recognition error increases (Fig. 2B). Figure 2B shows that the growth of the error is slowed down in the middle when the uncertainty unit has a value of 6. It is caused by the theta rhythm locking mechanism. When the uncertainty of the next phoneme is small, context is easily formed and used for predictions. In this case, theta rhythm locking is not effective according to the energy minimization hypothesis [34, 50] (Fig. 2C). As the uncertainty increases, phonemes from the spectrogram need to be decoded more reliably, and the theta coupling with the speech envelope becomes stronger. This enables speech recognition even in conditions of great uncertainty. An erroneous context prediction leads to a mismatch between the perception and the prediction processes. To rectify the error and select a new context in the model, information about the mismatch is relayed back up the hierarchy. The error is detected as a surprise at the top level (Fig. 2E), and this increases noise amplitude at the top level in order to drive explorative context switching. This increase in activity at the top of the model hierarchy is considered to be similar to error-related potentials in associative auditory cortical activity [34, 48].

BRyBI is largely invariant to the speaker's voice characteristics, e.g., due to speaker gender or dialect (Fig. 3A,B). The model performance follows observed data in recognition scores with speech rates [3, 4]; we see a relative robustness to speech rate until a critical compression ratio, beyond which performance degrades linearly (Fig. 3C). Performance is hypothesized to drop because accelerating speech faster than twice leads to a critical reduction in the length of the processing windows for syllables [4, 5] (see Table S1 for syllable and phrase duration statistics). Under this hypothesis, syllables that alternate faster than the theta rhythm cannot be parsed. This hypothesis has been tested by examining the limitations of human speech perception of interrupted as well as compressed and rechunked speech [1, 2, 5]. We thus set out to expose our model to such modulated speech signals to examine how rhythm-modulation of the generative inference process may account for the limitations of human perception.

First, we stimulated the model with speech that is interrupted by silent deletions. We find that the BRyBI model successfully accounts for the behavior in tasks with interrupted speech [1]. Here, normal-speed speech was cut in by a silent interval of various durations, during which speech information was lost (Fig. 3D, left). As in the experiments, the BRyBI showed a drop in the articulation score in performance at

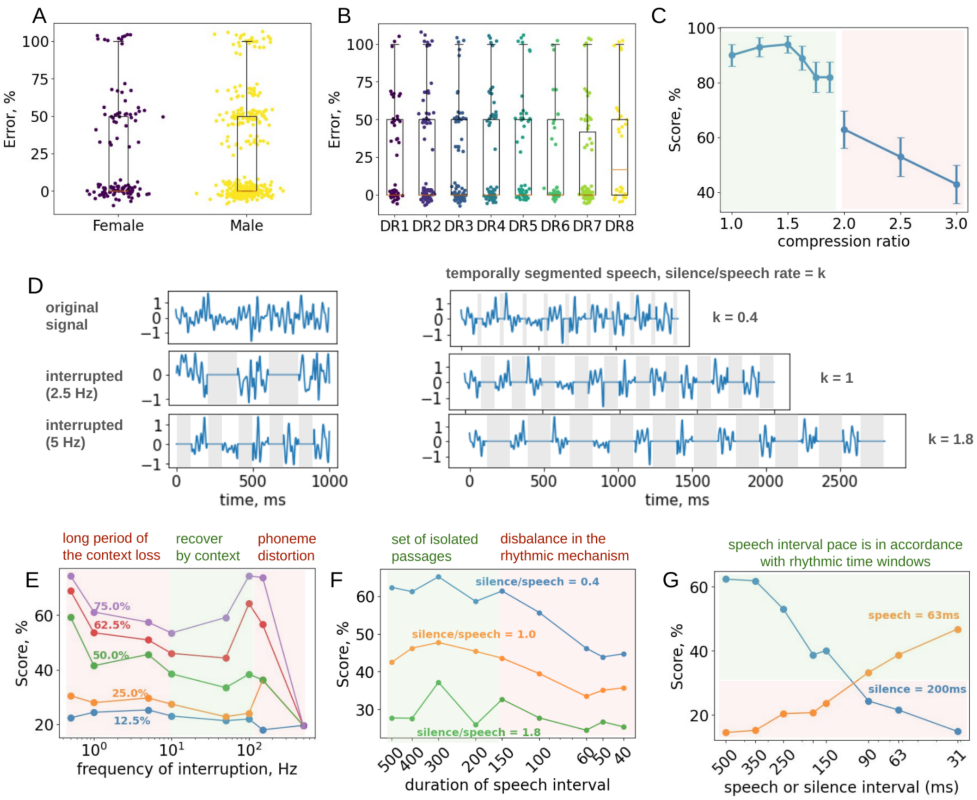


**Fig. 3** BRyBI performance for invariant speech parsing. Invariant performance in conditions of different speaker genders (A), dialects DR1-DR8 in the TIMIT dataset (B), and speech rates from 1 to 3 (C). When the speech rate is below 2, intelligibility remains consistently high, aligning with an average syllable frequency of less than 10 Hz and an average phrase frequency of less than 4 Hz (Tab. S1). However, a compression ratio exceeding 2 results in syllables and phrases extending beyond the limits of theta and delta rhythms, respectively. (D) Signal processing for the experiment with interrupted speech is a convolution of the original signal with a rectangular signal. Signal processing for temporally segmented speech includes silence insertion of different duration. On the right plots there are different silence-to-speech rates for fixed duration of speech intervals. (E) Experiment with interrupted speech. Different lines correspond to different speech-to-silence rates. A drop in the score at the 1 Hz interruption is indicative of prolonged information loss. A peak between 10 and 100 Hz represents the optimal interruption rate for context to recover missed information. Conversely, a decline at interruption frequencies exceeding 100 Hz is linked to impaired phoneme decoding from a distorted spectrogram. (F, G) Experiment with temporally segmented speech. (F) Isolated context units, represented by speech segments separated by long silent intervals, yield relatively high speech understanding. The score decreases as silent intervals shorten. We tested various silence-to-speech rates: 0.4, 1.0, and 1.8. BRyBI struggled most at a rate of 1.8. (G) BRyBI generates a cross-shaped plot for speech comprehension, dependent on speech and silent interval durations. Silent intervals at 200 msec, while varying speech intervals result in a decline in intelligibility (blue line). Similarly, fixing speech-interval duration at 63 msec and increasing silent-interval duration leads to decreased intelligibility (orange line).

the interruption frequency of 1 Hz that is caused by a long information loss interval (Fig. 3E, Fig. S3A). We also saw peaks in performance between 10 and 100 Hz,

which is the optimal interruption rate where context can recover missed information (Fig. 3E, Fig. S3B). Finally, the drop at the interruption frequency higher than 100 Hz is provoked by an impairment of phoneme decoding from a distorted spectrogram (Fig. 3E, Fig. S3C).

We then examined how the BRyBI model would perform under speech that was temporally segmented. Here, speech was not interrupted by silent gaps, but interspaced by added silent segments (Fig. 3D, right). Thus, there was no loss of information but a change in the timing of chunks' presentation. The model exhibits a behavior coherent with the human behavior with such temporally segmented speech [2] (Fig. 3F,G). Speech segments separated by long silent intervals can be considered isolated context units (a plateau for long chunks in Fig. 3F, Fig. S4A), resulting in a relatively high understanding of speech, which decreases as silent intervals become shorter (the decline for shorter intervals is shown in Fig. 3F, Fig. S4B). We tested conditions with different silence-to-speech rates (Fig. 3F), and for a value of 1.8, similar to the experiments [2], the BRyBI model had its lowest performance. Furthermore, BRyBI shows the same cross-shaped plot as in [2] for speech comprehension, depending on the duration of speech and silent intervals (Fig. 3G, Fig. S5). Maintaining silent intervals at a consistent 200 msec while varying speech intervals reveals a decline in intelligibility for speech intervals ranging from 200 to 31 msec (Fig. 3G, blue line). Similarly, when speech-interval duration is fixed at approximately 63 msec, increasing silent-interval duration from 63 to 500 msec results in a decrease in intelligibility (Fig. 3G, orange line). This aligns well with findings from the experimental study by Huggins et al. [2]. The experiment underscores that the intelligibility of temporally segmented speech hinges on the combined durations of speech and silent intervals. Reproducing these outcomes, BRyBI offers insights into the potential mechanisms at play, that is, a contextual support whose formation and transmission are controlled by the delta rhythm.

We next turned to an experiment that we reasoned would allow us to test our main mechanistic hypothesis: that the pattern of invariant speech recovery seen in humans is critically dependent on the delta-modulated top-down inference of the semantic context. In this experiment, speech was modulated by a combination of speech compression and temporal segmentation [5]. The duration of the silences inserted between segments varied between 0 and 160 msecs. The resulting errors showed a characteristic U-shaped plot of errors in speech recognition depending on the insertion.

Figure 4 shows an example of the full BRyBI model performance. In the control sentence, where no preprocessing was applied, BRyBI correctly reconstructs speech (15% of word errors for 100 sentences). For compressed speech, the model parallels a drop in human behavioral performance. Here, speech compression drives the frequency of phrases in the sentence beyond the delta range. This in turn prevents the model from providing the correct context to help with speech parsing (55% of word errors for 100 sentences). Respacing the syllables with silences recovers the speech rhythm, and thus leads to an improvement in speech intelligibility (Fig. 4A, Fig. 4C), reproducing the experimental U-shaped curve [5]. The previous hypothesis states that the insertion of silence intervals restores the syllabic rhythm, thereby restoring speech perception. In fact, the insertion of silence between syllables does not necessarily support

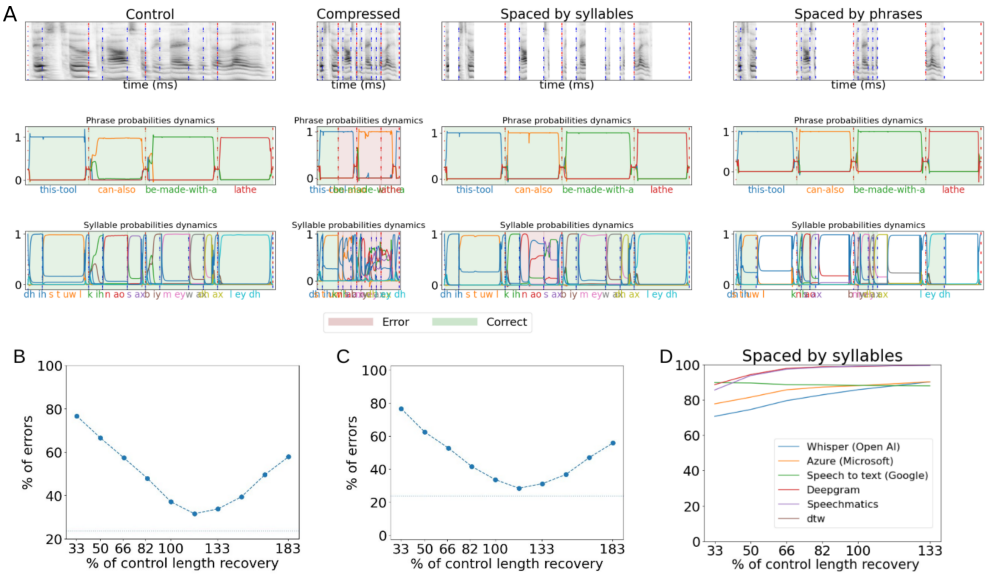


**Fig. 4** (A) Example of speech recognition in a test sentence: "this tool can also be made with a lathe '. Correct/incorrect recognition is indicated by the green/red background, respectively. Top panel: spectrogram; middle: reconstructed contextual candidates; bottom: reconstructed syllables; Note the errors in syllable and word sequencing. "Spaced by syllables" for inter-syllable 100 ms silence inserts; "Spaced by prosodic phrases" for inter-phrase 300 ms silence inserts. The control sentence, with no preprocessing, is reconstructed correctly. Compression shortens the time window for context formation and alternation, thus causing more errors. Respacing syllables and prosodic phrases by inserting silent intervals restores speech rhythm and improves intelligibility. (B, C) Speech recognition by BRyBI for temporally segmented compressed speech. Simulations do not only reproduce the U-shaped curve in the Ghitza and Grinberg 2009 experiment [5] but also predict the same error pattern when silent intervals are inserted between prosodic phrases. (D) Performance of AI models for speech recognition. All models exhibit low error rates in natural speech recognition (2.9–5.7%, Table S1) but face significant degradation in 3-time compressed speech (67.9–99.1%, Table S1). Segmentation of speech by silent intervals does not enhance comprehension for any of the AI models. The difference between AI models and human speech recognition suggests that hierarchical syntactic processing alone is not enough; the critical missing element is the temporal structuring that BRyBI implements utilizing brain rhythms.

this hypothesis, because speech perception may be restored due to the restoration of the rhythm of phrases rather than that of syllables. This means that if there is silence between syllables, not only the rhythm of the syllables, but also that of phrases and words is retrieved. At this point the model results deviate from the previous hypothesis and predict that restoring the rhythm of phrases (even without restoring that of syllables, when silence is inserted only between phrases) enables speech perception restoration. The following experiment further supports this point. Figure 4A and Figure 4C show that phrase recognition errors are bound to decrease when the natural rhythm of prosody is restored (Fig. 4B). BRyBI simulations thus suggest that the delta rhythm predictively sets the temporal boundaries for speech integration, beyond which speech becomes illegible.

To demonstrate that the delta-rhythm-modulated context inference is necessary to explain the brain mechanisms of human speech recognition patterns under modulated speech, we subjected BRyBI to two ablations: a context-free model (Fig. S1A) without the top layer and an arbitrary-timed context model (Fig. S1B) where the top-layer was not rhythmically modulated. Neither the context-free model (Fig. S1A) that only implements the bottom level, theta-gamma syllable code, nor the arbitrary-timing context model, where top-level predictions can switch at any time, reproduces the experimental results (Figures S1A,B). Interestingly, while the arbitrary context model (Fig. S1B, bottom panel) does not produce the telltale U-shape performance, it clearly demonstrates the benefits of context support since it yields low errors across all experimental conditions. The experimentally observed U-shaped error dependency on silence duration is reproduced by adding the delta-band temporal windowing for context alteration in the full model model (Fig. 1A and Fig. S1C).

In order to further show the key role of rhythmically modulated active interference in brain speech processing, we compared BRyBI performance with several LLMs for speech recognition where, to the best of our knowledge, such temporal processes are absent: Whisper (OpenAI) [72], Speech to text (Microsoft Azure), Speech to text (Google), Deepgram, and Speechmatics. These models use the hierarchy of speech organization and Bayesian inference, yet they do not include any notion of temporal windowing for the inference process. All models show low error rates in word recognition for natural speech (2.9 - 5.7%, Table S1) and significant degradation for compressed speech over two times (67.9 - 99.1%, Table S1). Respacing the speech chunks (syllables or context units) by silences had no effect on improvement in speech comprehension (Fig. 4D). From these simulations, we may speculate that even though LLMs show high performance for natural speech, the mechanisms for invariant speech recognition in LLMs and the human brain differ. Notably, hierarchical syntactic processing alone is insufficient to allow behavioral experiments with temporally segmented compressed speech to be reproducible. We believe that the key missing element is the temporal structuring of the predictive coding of speech information by endogenous brain rhythms.

# 3 Discussion

In this work, we provide a computational framework for understanding the role of brain rhythms in predictive coding and highlight the importance of uncertainty and surprise in this process. We show how an inferential theta-gamma code, together with the descending predictive influence of delta rhythm, converse in a predictive generative entrance model to produce precise and efficient neural processing of speech. Our model is able to address a three-part challenge. We show that BryBI can match patterns of robust and invariant speech processing that are seen in human experiments. Second, we suggest that the model does so in a biologically plausible manner, incorporating several mechanisms crucial to audio processing during speech perception. More specifically, within a predictive coding model, we implemented a mechanistically plausible hierarchical structure for syntax processing [6, 13] and oscillatory activity [17]. We note that predictive coding has recently received substantial support as a plausible

framework for speech processing [49, 58]. Finally, by modeling the rhythm generation and content formation phenomenologically, we were able to achieve the results with a relatively low model complexity, despite its robustness to extraneous perturbations of speech. This relatively low model complexity contrasts with the high complexity of the prevalent AI models of speech recognition and their need for large-scale computing resources, training data requirements, and energy consumption [73]. We note that despite the phenomenological form of the model equations, the inherent model structure emphasizes the biological plausibility of its component processes (rhythms, representation formation, theta-gamma encoding of syllables).

A compelling hypothesis posits that the core mechanism underpinning speech processing within the auditory cortex involves resolving a two-component optimization task within the framework of predictive coding — minimization of uncertainty and surprise [34, 49, 74]. Uncertainty and surprise in speech signals highlight the dynamic and predictable nature of phoneme transitions, with uncertainty reflecting a general lack of confidence in predicting the next phoneme and surprise denoting the occurrence of an unexpected phoneme in the input. Recently, a study combining non-invasive imaging and computational modeling with deep neural networks [49] demonstrated that the reduction of uncertainty in words can be explained by an increase in surprise (thus, updates in GM) and correlates with delta rhythm in aAC, while the uncertainty of phonemes correlates with modulation of theta rhythm in pAC. The BRyBI model shows how this hypothesis can be substantiated computationally through a synergy of predictive coding and oscillatory activity with only a minimum number of layers. Replicating the processes between the midbrain and the pAC, the bottom level is designed to receive sensory input information, including spectrograms and modulation signals. Using this bottom-up information, the model efficiently disentangles phonemes and syllables by minimizing phonemic uncertainty in a way that is compatible with several recently proposed feedforward models [5, 28, 29, 53]. In BRyBI, a delta-modulated top-down semantic context inference process guides this bottom-up information flow.

The present computational work shows that top-down context inference and its governance by the delta-rhythm is critical to account for the patterns of human speech processing. In fact, when considering only feed-forward processes, the human performance patterns cannot be reproduced (i.e. the purely feed-forward context-free model fails (Fig. S1A)). The experiments that we address specifically tested for invariance to degradation, segmentation, and the recovery of performance under re-spacing of audio signals. Critically, in the model allowing for the descending context-formation process to run unfettered by a brain rhythm, we do not reproduce human-like performance, showing that this degraded model is insensitive to speech manipulations (Fig. S1). On the other hand, the rhythmically-governed context formation model accounted for both the patterns of invariance and the recovery of distorted speech.

BRyBI allows us to go beyond just reproducing the phenomenology of behavior but to understand in detail how the predictive coding computations combine with oscillatory temporal governance to orchestrate the necessary brain computations. If we track the computational process within our model, we see that at each time step, the feed-forward module reproduces the dynamics of syllables predetermined for each phrase in the correct order, as captured and governed by the delta module. When a

critical discrepancy arises between the generated theta-gamma code and the incoming sensory input, the encoding process at the bottom level deviates from the predicted context. This conflict, in turn, triggers an update of contextual beliefs at the top level. This update induces a sharp shift in the dynamic state of the context module, requiring a transient increase in this module activity (see details in the section: Methods). In essence, consistent with previous findings in [49, 75], the reaction to surprise increases delta rhythm activity. Interestingly, this increase in corrective activity reproduces the phenomenology of the ERP signal [34, 39, 76], offering insights into the hypothetical mechanisms underlying ERP in the brain.

Specifically, minimizing surprise during online speech recognition is associated with the goal of selecting the most appropriate context. The judicious choice of context allows the theta rhythm not to be perfectly synchronized with speech, according to the energy minimization hypothesis [34, 50]. Should the context be erroneous, the model effectively needs to update its corresponding state (i.e., the context of a phrase or word). Such a switch requires a time-locked increase in activity at the context level. We can surmise that such an increase underpins the error-related potentials seen during complex speech recognition tasks [48, 61]. As a result, the BRyBI model lends support to the hypothesis that top-down predictive and bottom-up acoustic flows are dynamically integrated, as proposed by several studies [38, 77, 78].

Recent studies show a structural hierarchy in the processing of several speech features and highlight the relationship between this hierarchy and the organization of rhythmic activity [63]. In particular, it has been discovered that the theta rhythm's entrainment is correlated with speech clarity and acoustic properties, whereas the delta rhythm's entrainment is correlated with higher-order speech comprehension. Based on these findings, BRyBI suggests processing syntactic speech units in a sequential manner, with phonemes and syllables processed at the bottom level and words and phrases processed at the top level. In particular, phonemes and syllables are associated with "fast" gamma and theta rhythms, respectively, whereas words and phrases are associated with the slow delta rhythm. Within BRyBI, the level of syllables integrates information from both the bottom-up acoustic signal and the top-down contextual signal.

One of the strongest arguments for questioning rhythm-based speech parsing is that theta-locking can vary significantly across different experiments. For example, [23] and [4] showed that theta was strongly locked, while other experiments found weak locking despite good behavior performance [43, 79]. We can propose an explanation for this ambiguous evidence using rhythm-based predictive coding for speech recognition. According to our model, theta-locking is flexible: in clear contexts, it is floating; in unclear contexts, the speech envelope needs to entrain the theta rhythm [28, 80].

Although BRyBI shows promising results, it leads to multiple avenues for extensions and improvements through the implementation of more biological mechanisms for rhythm generation, the incorporation of phase-amplitude coupling (PAC) mechanisms, and considering the role of beta in the inference hierarchy [81].

Another future direction can be to expand and improve the linguistic part of the model. For example, several models that propose the incorporation of compositional mechanisms [74, 82–85] can extend the BRyBI model for the semantic part. Some of

these models [74, 82] illustrate language representations processing using asynchrony and inhibition. A biophysical version of BRyBI, e.g., where rhythms are implemented with dynamical neural mass- or spiking-networks [17, 39, 86, 87], could usefully integrate these concepts and mechanisms. This would allow for direct comparisons with electrophysiological experimental data that may aid not only in data-based model identification but also reveal the fundamental theory of speech coding in the brain.

The practical implications of studies on neural oscillations and their ability to partly synchronize with external stimuli are important for the treatment of a variety of pathologies. For example, an experiment found a relationship between the synchronization of delta- and gamma-band networks and semantic fluency in post-stroke chronic aphasia [88]. Similarly, a study found tracking of theta rhythms but not delta rhythms in the logopenic variant of primary progressive aphasia, which may indicate ineffective top-down coding [89]. Dyslexia is another pathology that has been associated with disturbances in low-frequency rhythm tracking [90–94]. According to the rise-time theory of dyslexia, reading difficulties result from the complexity of tracking the amplitude modulation of external signals, which leads to difficulties in speech perception, phonological processing, and, ultimately, reading [94]. For instance, children with dyslexia have abnormal delta rhythm phase alignment when interpreting rhythmic syllable sequences, which affects speech representation [95]. Likewise, children with dyslexia showed impaired tapping to a metronome beat with a frequency of 2 Hz [96]. Another study compared neural responses to speech and non-speech sounds in healthy people and those with dyslexia, revealing that healthy people had stronger delta-band responses in the right hemisphere and gamma-band responses in the left hemisphere [97]. In these experiments, difficulties in tracking low-frequency external rhythms were correlated with phonetic perception problems. It is noteworthy that individuals with dyslexia can compensate for their phonological perception deficits with semantic context, i.e., top-down compensatory mechanisms [98–100].

Taken together, these findings suggest that disturbances in low-frequency rhythm synchronization may be a factor that affects the progression of aphasia and dyslexia. Thus, the success of transcranial electrical stimulation in treating these conditions may be partially explained by the synchronization of neural oscillations with external stimuli [101–103]. Clearly, the practical application of such research must rely heavily on a solid theoretical framework capable of predicting treatment effects, developing hypotheses, and developing experimental and treatment protocols. The BRyBI model can provide such a theoretical basis.

In summary, our results shed light on the intrinsic constraints and compensatory mechanisms of human speech perception. At the same time they offer a potential alternative and challenge to the prevalent AI NLP approaches to speech processing, pointing out how brain mechanisms may allow for robust speech recognition with high computational efficiency, even under conditions where LLMs appear to perform poorly.

# 4 Methods

Predictive coding implies two directions of information flow: top-down and bottom-up. Top-down flow is provided by constructing the generative model (GM) and passing

beliefs from higher abstract levels to the early sensory areas. Bottom-up flow propagates updates in beliefs during the inference process provided by the DEM algorithm [57].

GM is essentially a stochastic dynamical system that has a hierarchical structure. Each level is formed by two types of variables: hidden and causal states. The hidden states are ruled by differential equations. The causal states serve to transfer information from the top to the bottom levels and represent beliefs inferred from the internal model of the world. They are formed as nonlinear transformations of the hidden states. Thus, GM maintains a top-down information flow. The BRyBI model consists of two levels and is formalized as follows:

$$\begin{cases} \dot{x}^{(2)} = f^{(2)}(x^{(2)}) + \epsilon^{(2)}, \\ \nu^{(2)} = g^{(2)}(x^{(2)}) + \eta^{(2)}, \\ \dot{x}^{(1)} = f^{(1)}(x^{(1)}, \nu^{(2)}) + \epsilon^{(1)}, \\ \nu^{(1)} = g^{(1)}(x^{(1)}, \nu^{(2)}) + \eta^{(1)}. \end{cases} \tag{1}$$

Here $x^{(i)}$ is the hidden state of $i$-level with a noise $\epsilon^{(i)}$, $\nu^{(i)}$ is the corresponding causal state with a noise $\eta^{(i)}$. The function $f^{(i)}$ determines a form of differential equations for the hidden state $x^{(i)}$. The function $g^{(i)}$ determines the nonlinear transformation of $x^{(i)}$ taking into account information from the level above.

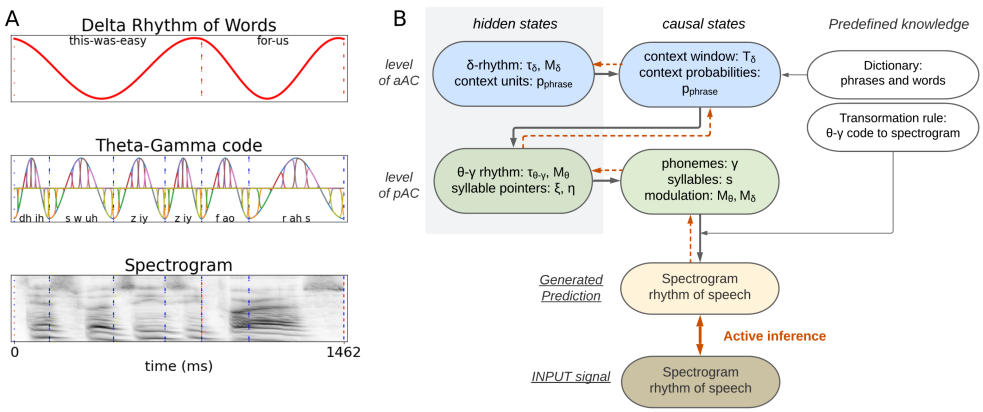


**Fig. 5** (A) An example of rhythmic activity reconstructed by the model: delta rhythm (top), coupled theta-gamma rhythm (middle), and ground-truth spectrum (bottom). Red dashed lines are ground-truth word boundaries; blue dashed lines are ground-truth syllable boundaries. See the text for the model's notation. (B) A detailed structure of BRyBI.

The bottom and top levels in BRyBI model mimic speech processing in the primary auditory cortex (pAC) and the associative auditory cortex (aAC), respectively (Fig. 5B). On the top level delta rhythm switches words. On the bottom level, coupled theta and gamma rhythms code acoustic signal of speech.

## 4.1 The top level

The top level is the level of context candidates that are alternating in delta rhythm. Here we define a delta-timescale:

$$\begin{cases} \dot{\tau}_\delta = \ln(1 + \exp(M_\delta)) + \epsilon^{(2)}_{\tau_\delta}, \\ \dot{M}_\delta = -k_\delta(M_\delta - M_0)^3 + \epsilon^{(2)}_{M_\delta}, \end{cases} \tag{2}$$

where the delta-timescale $\tau_\delta$ is phase modulated by $M_\delta$. The parameter $k_\delta$ constrains the timescale in the delta band. $M_\delta$ potentially takes values from $[-\infty, \infty]$. The function $\text{softplus}(M_\delta) = \ln(1 + \exp(M_\delta))$ maps the values to $[0, +\infty]$. When $M_\delta = M_0$ and the delta-wave does indeed have an average delta-rhythm frequency. In order to change this state, it is necessary to increase / decrease the noise $\epsilon^{M_\delta}$. Thus, the more the frequency differs from the delta rhythm, the more difficult it is to obtain the corresponding $M_\delta$ at the expense of noise.

Delta-waves are constructed as follows:

$$\delta_{wave_i} = \cos(v_\delta \cdot \tau_\delta - \frac{2\pi i}{100}), \tag{3}$$

where $v_\delta = \frac{2\pi\Omega_\delta}{1000}$, $\Omega_\delta = 2.9$ Hz is the average frequency of delta rhythm.

Phrases are chosen from the language randomly. The relative probability of each phrase accumulates in the variable $\mathbf{w}$:

$$\frac{d\mathbf{w}}{dt} = -\mathbf{w}T_\delta + \epsilon^{(2)}_w, \tag{4}$$

On a certain phase of the delta rhythm trigger, $T_\delta = \text{softmax}(\delta_{waves})$ switches phrases by abruptly increasing from 0 to 1.

On this level, hidden states are ($\tau_\delta$, $M_\delta$) and $\mathbf{w}$. Causal states are $\nu^{(2)}_{\tau_\delta} = \tau_\delta + \eta^{(2)}_{\tau\delta}$, $\nu^{(2)}_{M_\delta} = M_\delta + \eta^{(2)}_{M\delta}$ and word probabilities $\nu^{(2)}_w = \text{softmax}(\mathbf{w}) + \eta^{(2)}_w$.

## 4.2 The bottom level

The theta-timescale defines a window of syllable coding. We use the same model as for the delta-timescale:

$$\begin{cases} \dot{\tau}_\theta = \ln(1 + \exp(M_\theta)) + \epsilon^{(1)}_{\tau_\theta}, \\ \dot{M}_\theta = 0 + \epsilon^{(1)}_{M_\theta}. \end{cases} \tag{5}$$

Following the example of previous similar models [29, 60, 81], the GM splits each syllable into 8 parts. It allows more flexibility in shaping the auditory spectrogram of syllables and phonemes. The gamma waves are constructed as a nonlinear function of $\tau_\theta$ and each gamma wave has a period equal to 1/8 of the period of a syllable:

$$\gamma_i = \text{softmax}(30 \cdot \sin(2\pi(\tau_\theta - \phi_i))), \tag{6}$$

where $\phi = i/8$, indexes $i = 0..7$.

Syllable selection is a crucial module in speech interpretation. In BRyBI, we want to enter the context of a certain phrase for its constitutive syllables. A phrase is basically

an ordered sequence of syllables. In this definition, it is convenient to represent it as a matrix of syllables (Fig. 6):

$$W_k = ||W_{ij}||_{i=1..m}^{j=1..m}, \tag{7}$$

where $m$ is a whole number of syllables in the language. An element $W_{ij} = 1$ if $j$-th syllable follows the $i$-th syllable; otherwise, $W_{ij} = 0$.

The context matrix is defined as the weighted sum of matrices of syllables: $W = \sum_{k=1}^{n} \nu_w^{(2)} W_k$, $n$ is a number of words and phrases in the dictionary. If exactly one word were chosen in the variable $\mathbf{w}$, i.e., only one value in the vector was equal to 1, and all the rest were equal to zero, then such a sum would choose from all matrices $W$ only the one corresponding to the current word. The matrix $W$ changes dynamically depending on the word/phrase probabilities at the top level.

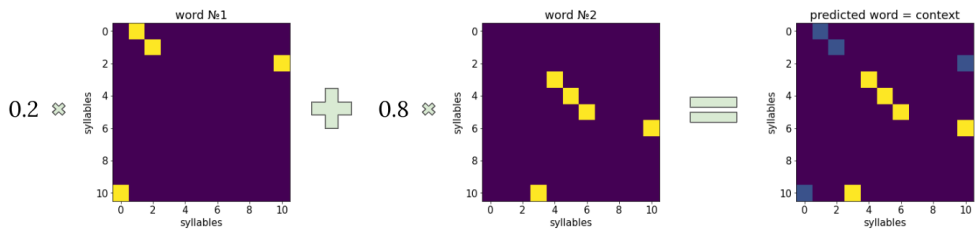


**Fig. 6** Example of word representations and context construction as the sum of word matrices normalized by probabilities from the top level.

Syllable transition frequency inside a word / phrase is in the theta-band, whereas word switching frequency is in the delta-band. Hidden states of syllables are determined as follows:

$$\begin{cases} \dot{\boldsymbol{\eta}} = \kappa_s(W\boldsymbol{\xi} - \boldsymbol{\eta})(1-c) - (\boldsymbol{\eta} - \boldsymbol{\eta}_{\#}) \cdot T_\delta + \epsilon_\eta^{(1)}, \\ \dot{\boldsymbol{\xi}} = \kappa_s(W\boldsymbol{\eta} - \boldsymbol{\xi})c - (\boldsymbol{\xi} - \boldsymbol{\xi}_{\#}) \cdot T_\delta + \epsilon_\xi^{(1)}, \\ c = \frac{\exp(\sin^2(\tau_\theta))}{\exp(\sin^2(\tau_\theta)) + \exp(\cos^2(\tau_\theta))}, \\ \boldsymbol{s} = (1-c)\boldsymbol{\eta} + c\boldsymbol{\xi} \end{cases} \tag{8}$$

Here $\eta$ and $\xi$ are two syllable pointers. The equations determine a gradual transition from one syllable, that is pointed by $\xi$, to another syllable, that is pointed by $\eta$, with a speed $\kappa_s$. These two pointers essentially follow their own theta wave. At the same time, the expected syllable from the context is encoded in half of the cycle; in the second half, it either occurs or it can knock out another syllable by error (if the context, for example, was chosen incorrectly).

On the bottom level hidden states are the theta-timescale $\tau_\theta$, theta modulation $M_\theta$, and syllable pointers $\xi$ and $\eta$.

A theta-gamma code of an acoustic signal generates as a convolution of syllable and gamma-units with predefined for each frequency band $f$ tensors $P_{f\gamma\theta}$:

$$\boldsymbol{c} = \sum_{i=1}^{8} \sum_{j=1}^{m} \gamma_i s_j P_{fij} + \epsilon_c^{(1)}. \tag{9}$$

Generated auditory spectrogram $\mathbf{c}$, delta $M_\delta$ and theta $M_\theta$ modulations are compared with an input. The DEM-algorithm produces joint distributions for all hidden and causal variables, which are used in the model to recognize syllables, words and phrases.

### 4.3 The dataset

The extraction of syllable matrices is as follows. For each sentence, for each syllable in the sentence:

1. A piece of the spectrum is extracted according to the boundaries of the syllable;
2. A piece of the spectrum is split into 8 equal segments;
3. Over time, each part is averaged. The result is eight 6-dimensional vectors, one for each scale.

**Acknowledgments.** This publication is supported by the Brain Program of the IDEAS Research Center and the Vernadski scholarship. The research in part through computational resources of HPC facilities at HSE University. BSG was supported by CNRS, INSERM.

## References

[1] Miller, G.A., Licklider, J.C.: The intelligibility of interrupted speech. The Journal of the Acoustical Society of America **22**(2), 167–173 (1950)

[2] Huggins, A.: Temporally segmented speech. Perception & Psychophysics **18**, 149–157 (1975)

[3] Garvey, W.D.: The intelligibility of speeded speech. Journal of experimental psychology **45**(2), 102 (1953)

[4] Ghitza, O.: Behavioral evidence for the role of cortical $\theta$ oscillations in determining auditory channel capacity for speech. Frontiers in psychology **5**, 652 (2014)

[5] Ghitza, O., Greenberg, S.: 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 (2009)

[6] Stephenson, C., Feather, J., Padhy, S., Elibol, O., Tang, H., McDermott, J., Chung, S.: Untangling in invariant speech recognition. Advances in neural information processing systems **32** (2019)

[7] Greenberg, S., Kingsbury, B.E.: The modulation spectrogram: In pursuit of an invariant representation of speech. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1647–1650 (1997). IEEE

[8] Kösem, A., Bosker, H.R., Meyer, A.S., Jensen, O., Hagoort, P.: Neural entrainment reflects temporal predictions guiding speech comprehension. In: The Eighth Annual Meeting of the Society for the Neurobiology of Language (snl 2016) (2016)

[9] Kösem, A., Bosker, H.R., Takashima, A., Meyer, A., Jensen, O., Hagoort, P.: Neural entrainment determines the words we hear. Current Biology **28**(18), 2867–2875 (2018)

[10] Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J.T., Hickok, G.: Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cerebral Cortex **20**(10), 2486–2495 (2010)

[11] Evans, S., Davis, M.H.: Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cerebral cortex **25**(12), 4772–4788 (2015)

[12] Obleser, J., Leaver, A.M., VanMeter, J., Rauschecker, J.P.: Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Frontiers in psychology **1**, 232 (2010)

[13] Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E.: The hierarchical cortical organization of human speech processing. Journal of Neuroscience **37**(27), 6539–6557 (2017)

[14] Caucheteux, C., Gramfort, A., King, J.-R.: Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour, 1–12 (2023)

[15] Longacre, R.E.: Hierarchy in language. Method and theory in linguistics, 173–195 (1970)

[16] Poeppel, D.: The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. Cognitive neuropsychology **29**(1-2), 34–55 (2012)

[17] Giraud, A.-L., Poeppel, D.: Cortical oscillations and speech processing: emerging computational principles and operations. Nature neuroscience **15**(4), 511–517 (2012)

[18] Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D.: Cortical tracking of hierarchical linguistic structures in connected speech. Nature neuroscience **19**(1),

158–164 (2016)

[19] Ronconi, L., Oosterhof, N.N., Bonmassar, C., Melcher, D.: Multiple oscillatory rhythms determine the temporal organization of perception. Proceedings of the National Academy of Sciences **114**(51), 13435–13440 (2017)

[20] Chen, B., Ciria, L.F., Hu, C., Ivanov, P.C.: Ensemble of coupling forms and networks among brain rhythms as function of states and cognition. Communications Biology **5**(1), 82 (2022)

[21] Buzsáki, G., Watson, B.O.: Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease. Dialogues in clinical neuroscience (2022)

[22] Hyafil, A., Giraud, A.-L., Fontolan, L., Gutkin, B.: Neural cross-frequency coupling: connecting architectures, mechanisms, and functions. Trends in neurosciences **38**(11), 725–740 (2015)

[23] Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S.: Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS biology **11**(12), 1001752 (2013)

[24] Aiken, S.J., Picton, T.W.: Human cortical responses to the speech envelope. Ear and hearing **29**(2), 139–157 (2008)

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

[26] Tang, C., Hamilton, L., Chang, E.: Intonational speech prosody encoding in the human auditory cortex. Science **357**(6353), 797–801 (2017)

[27] Murphy, E.: A theta-gamma neural code for feature set composition with phase-entrained delta nestings. UCL Work. Pap. Linguist **28**, 1–23 (2016)

[28] Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., Giraud, A.-L.: Speech encoding by coupled cortical theta and gamma oscillations. Elife **4**, 06213 (2015)

[29] Hovsepyan, S., Olasagasti, I., Giraud, A.-L.: Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. Nature communications **11**(1), 1–12 (2020)

[30] Shamir, M., Ghitza, O., Epstein, S., Kopell, N.: Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. PLoS computational biology **5**(5), 1000370 (2009)

[31] Züst, H., Tschopp, K.: Influence of context on speech understanding ability using german sentence test materials. Scandinavian audiology **22**(4), 251–255 (1993)

[32] Golumbic, E.M.Z., Poeppel, D., Schroeder, C.E.: Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. Brain and language **122**(3), 151–161 (2012)

[33] Holt, L.L., Lotto, A.J.: Behavioral examinations of the level of auditory processing of speech context effects. Hearing research **167**(1-2), 156–169 (2002)

[34] Molinaro, N., Lizarazu, M., Baldin, V., Pérez-Navarro, J., Lallier, M., Ríos-López, P.: Speech-brain phase coupling is enhanced in low contextual semantic predictability conditions. Neuropsychologia **156**, 107830 (2021)

[35] Ten Oever, S., Carta, S., Kaufeld, G., Martin, A.E.: Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. Elife **11**, 77468 (2022)

[36] Herbst, S.K., Obleser, J.: Implicit temporal predictability enhances pitch discrimination sensitivity and biases the phase of delta oscillations in auditory cortex. NeuroImage **203**, 116198 (2019)

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

[38] Hannemann, R., Obleser, J., Eulitz, C.: Top-down knowledge supports the retrieval of lexical information from degraded speech. Brain research **1153**, 134–143 (2007)

[39] Forseth, K.J., Hickok, G., Rollo, P., Tandon, N.: Language prediction mechanisms in human auditory cortex. Nature communications **11**(1), 5240 (2020)

[40] Ding, R., Oever, S., Martin, A.E.: Pronoun resolution via reinstatement of referent-related activity in the delta band. bioRxiv, 2023–04 (2023)

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

[42] Myers, B.R., Lense, M.D., Gordon, R.L.: Pushing the envelope: Developments in neural entrainment to speech and the biological underpinnings of prosody perception. Brain sciences **9**(3), 70 (2019)

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

[44] Attaheri, A., Choisdealbha, Á.N., Di Liberto, G.M., Rocha, S., Brusini, P., Mead, N., Olawole-Scott, H., Boutris, P., Gibbon, S., Williams, I., *et al.*: Delta-and

theta-band cortical tracking and phase-amplitude coupling to sung speech by infants. NeuroImage **247**, 118698 (2022)

[45] Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., Morillon, B.: Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. PLoS biology **18**(3), 3000207 (2020)

[46] Ghitza, O.: Acoustic-driven delta rhythms as prosodic markers. Language, Cognition and Neuroscience **32**(5), 545–561 (2017)

[47] Rimmele, J.M., Poeppel, D., Ghitza, O.: Acoustically driven cortical $\delta$ oscillations underpin prosodic chunking. Eneuro **8**(4) (2021)

[48] Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., Haider, H.: Fractionating language comprehension via frequency characteristics of the human eeg. Neuroreport **15**(3), 409–412 (2004)

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

[50] Bai, F., Meyer, A.S., Martin, A.E.: Neural dynamics differentially encode phrases and sentences during spoken language comprehension. PLoS Biology **20**(7), 3001713 (2022)

[51] Park, H., Thut, G., Gross, J.: Predictive entrainment of natural speech through two fronto-motor top-down channels. Language, Cognition and Neuroscience **35**(6), 739–751 (2020)

[52] Meyer, L., Henry, M.J., Gaston, P., Schmuck, N., Friederici, A.D.: Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cerebral Cortex **27**(9), 4293–4302 (2017)

[53] Nabé, M., Schwartz, J.-L., Diard, J.: Cosmo-onset: A neurally-inspired computational model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets. Frontiers in Systems Neuroscience, 75 (2021)

[54] Ghitza, O.: On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Frontiers in psychology **3**, 238 (2012)

[55] Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience **2**(1), 79–87 (1999)

[56] Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J.: Canonical microcircuits for predictive coding. Neuron **76**(4), 695–711 (2012)

[57] Friston, K., Kiebel, S.: Predictive coding under the free-energy principle. Philosophical transactions of the Royal Society B: Biological sciences **364**(1521), 1211–1221 (2009)

[58] Heilbron, M., Chait, M.: Great expectations: is there evidence for predictive coding in auditory cortex? Neuroscience **389**, 54–73 (2018)

[59] Peelle, J.E., Johnsrude, I., Davis, M.H.: Hierarchical processing for speech in human auditory cortex and beyond. Frontiers in human neuroscience **4**, 51 (2010)

[60] Su, Y., MacGregor, L.J., Olasagasti, I., Giraud, A.-L.: A deep hierarchy of predictions enables online meaning extraction in a computational model of human speech comprehension. Plos Biology **21**(3), 3002046 (2023)

[61] Friston, K.J., Sajid, N., Quiroga-Martinez, D.R., Parr, T., Price, C.J., Holmes, E.: Active listening. Hearing research **399**, 107998 (2021)

[62] Friston, K.J., Parr, T., Yufik, Y., Sajid, N., Price, C.J., Holmes, E.: Generative models, linguistic communication and active inference. Neuroscience & Biobehavioral Reviews **118**, 42–64 (2020)

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

[64] Zhao, B., Dang, J., Zhang, G., Unoki, M.: Cortical oscillatory hierarchy for natural sentence processing. In: INTERSPEECH, pp. 125–129 (2020)

[65] Poeppel, D., Idsardi, W.J., Van Wassenhove, V.: Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1493), 1071–1086 (2008)

[66] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A.: The natural statistics of audiovisual speech. PLoS computational biology **5**(7), 1000436 (2009)

[67] Ghitza, O.: Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Frontiers in psychology **2**, 130 (2011)

[68] Rumelhart, D.E., McClelland, J.L.: An interactive activation model of context effects in letter perception: Ii. the contextual enhancement effect and some tests and extensions of the model. Psychological review **89**(1), 60 (1982)

[69] McClelland, J.L., Rumelhart, D.E.: An interactive activation model of context effects in letter perception: I. an account of basic findings. Psychological review

**88**(5), 375 (1981)

[70] Garofolo, J.S.: Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993 (1993)

[71] Tezcan, F., Weissbart, H., Martin, A.E.: A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. Elife **12**, 82386 (2023)

[72] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[73] Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., Poria, S.: A review of deep learning techniques for speech processing. Information Fusion, 101869 (2023)

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

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

[76] Lau, E.F., Phillips, C., Poeppel, D.: A cortical network for semantics:(de) constructing the n400. Nature reviews neuroscience **9**(12), 920–933 (2008)

[77] Ten Oever, S., Schroeder, C.E., Poeppel, D., Van Atteveldt, N., Zion-Golumbic, E.: Rhythmicity and cross-modal temporal cues facilitate detection. Neuropsychologia **63**, 43–50 (2014)

[78] Rimmele, J.M., Morillon, B., Poeppel, D., Arnal, L.H.: Proactive sensing of periodic and aperiodic auditory patterns. Trends in cognitive sciences **22**(10), 870–882 (2018)

[79] Canales-Johnson, A., Borges, A.F.T., Komatsu, M., Fujii, N., Fahrenfort, J.J., Miller, K.J., Noreika, V.: Broadband dynamics rather than frequency-specific rhythms underlie prediction error in the primate auditory cortex. Journal of Neuroscience **41**(45), 9374–9391 (2021)

[80] Giraud, A.-L.: Oscillations for all¯_ () _/¯? a commentary on meyer, sun & martin (2020). Language, Cognition and Neuroscience **35**(9), 1106–1113 (2020)

[81] Hovsepyan, S., Olasagasti, I., Giraud, A.-L.: Rhythmic modulation of prediction errors: A top-down gating role for the beta-range in speech processing. PLOS Computational Biology **19**(11), 1011595 (2023)

[82] Hummel, J.E., Holyoak, K.J.: Distributed representations of structure: A theory

of analogical access and mapping. Psychological review **104**(3), 427 (1997)

[83] Doumas, L.A., Hummel, J.E., Sandhofer, C.M.: A theory of the discovery and predication of relational concepts. Psychological review **115**(1), 1 (2008)

[84] Shastri, L.: Types and quantifiers in shruti–a connectionist model of rapid reasoning and relational processing. In: International Workshop on Hybrid Neural Systems, pp. 28–45 (1998). Springer

[85] Martin, A.E., Doumas, L.A.: A mechanism for the cortical computation of hierarchical linguistic structure. PLoS biology **15**(3), 2000663 (2017)

[86] Poeppel, D., Assaneo, M.F.: Speech rhythms and their neural foundations. Nature reviews neuroscience **21**(6), 322–334 (2020)

[87] Stanley, D.A., Falchier, A.Y., Pittman-Polletta, B.R., Lakatos, P., Whittington, M.A., Schroeder, C.E., Kopell, N.J.: Flexible reset and entrainment of delta oscillations in primate primary auditory cortex: modeling and experiment. BioRxiv, 812024 (2019)

[88] Mehraram, R., Kries, J., De Clercq, P., Vandermosten, M., Francart, T.: Eeg reveals brain network alterations in chronic aphasia during natural speech listening. bioRxiv, 2023–03 (2023)

[89] Dial, H.R., Gnanateja, G.N., Tessmer, R.S., Gorno-Tempini, M.L., Chandrasekaran, B., Henry, M.L.: Cortical tracking of the speech envelope in logopenic variant primary progressive aphasia. Frontiers in human neuroscience **14**, 597694 (2021)

[90] Lallier, M., Lizarazu, M., Molinaro, N., Bourguignon, M., Ríos-López, P., Carreiras, M.: From auditory rhythm processing to grapheme-to-phoneme conversion: How neural oscillations can shed light on developmental dyslexia. Reading and Dyslexia: From Basic Functions to Higher Order Cognition, 147–163 (2018)

[91] Di Liberto, G.M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., Lalor, E.C.: Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. NeuroImage **175**, 70–79 (2018)

[92] Power, A.J., Colling, L.J., Mead, N., Barnes, L., Goswami, U.: Neural encoding of the speech envelope by children with developmental dyslexia. Brain and Language **160**, 1–10 (2016)

[93] Destoky, F., Bertels, J., Niesen, M., Wens, V., Vander Ghinst, M., Rovai, A., Trotta, N., Lallier, M., De Tiège, X., Bourguignon, M.: The role of reading experience in atypical cortical tracking of speech and speech-in-noise in dyslexia. NeuroImage **253**, 119061 (2022)

[94] Goswami, U.: Sensory theories of developmental dyslexia: three challenges for research. Nature Reviews Neuroscience **16**(1), 43–54 (2015)

[95] Power, A.J., Mead, N., Barnes, L., Goswami, U.: Neural entrainment to rhythmic speech in children with developmental dyslexia. Frontiers in human neuroscience **7**, 777 (2013)

[96] Thomson, J.M., Goswami, U.: Rhythmic processing in children with developmental dyslexia: auditory and motor rhythms link to reading and spelling. Journal of Physiology-Paris **102**(1-3), 120–129 (2008)

[97] Lizarazu, M., Covella, L.S., Wassenhove, V., Rivière, D., Mizzi, R., Lehongre, K., Hertz-Pannier, L., Ramus, F.: Neural entrainment to speech and nonspeech in dyslexia: conceptual replication and extension of previous investigations. Cortex **137**, 160–178 (2021)

[98] Klimovich-Gray, A., Di Liberto, G., Amoruso, L., Barrena, A., Agirre, E., Molinaro, N.: Increased top-down semantic processing in natural speech linked to better reading in dyslexia. NeuroImage **273**, 120072 (2023)

[99] Giraud, A.-L., Ramus, F.: Neurogenetics and auditory processing in developmental dyslexia. Current opinion in neurobiology **23**(1), 37–42 (2013)

[100] Lehongre, K., Morillon, B., Giraud, A.-L., Ramus, F.: Impaired auditory sampling in dyslexia: further evidence from combined fmri and eeg. Frontiers in human neuroscience **7**, 454 (2013)

[101] Elsner, B., Kugler, J., Pohl, M., Mehrholz, J.: Transcranial direct current stimulation (tdcs) for improving aphasia in adults with aphasia after stroke. Cochrane Database of Systematic Reviews (5) (2019)

[102] Biou, E., Cassoudesalle, H., Cogné, M., Sibon, I., De Gabory, I., Dehail, P., Aupy, J., Glize, B.: Transcranial direct current stimulation in post-stroke aphasia rehabilitation: A systematic review. Annals of physical and rehabilitation medicine **62**(2), 104–121 (2019)

[103] Xie, X., Hu, P., Tian, Y., Wang, K., Bai, T.: Transcranial alternating current stimulation enhances speech comprehension in chronic post-stroke aphasia patients: A single-blind sham-controlled study. Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation **15**(6), 1538–1540 (2022)