



AFRL-AFOSR-VA-TR-2021-0068

Parsing continuous speech: the role of neuronal oscillations and sentential context

Ghitza, Oded
TRUSTEES OF BOSTON UNIVERSITY
1 SILBER WAY
BOSTON, MA, 2215
USA

07/07/2021
Final Technical Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 07-07-2021		2. REPORT TYPE Final		3. DATES COVERED (From - To) 01 Dec 2017 - 31 Mar 2021		
4. TITLE AND SUBTITLE Parsing continuous speech: the role of neuronal oscillations and sentential context				5a. CONTRACT NUMBER		
				5b. GRANT NUMBER FA9550-18-1-0055		
				5c. PROGRAM ELEMENT NUMBER 61102F		
6. AUTHOR(S) Oded Ghitza				5d. PROJECT NUMBER		
				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) TRUSTEES OF BOSTON UNIVERSITY 1 SILBER WAY BOSTON, MA 2215 USA				8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AF Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203				10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/AFOSR RTB2		
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-AFOSR-VA-TR-2021-0068		
12. DISTRIBUTION/AVAILABILITY STATEMENT A Distribution Unlimited: PB Public Release						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT The objective of this research program is to formulate a fundamental cortical computation principle, universal across sensory modalities, which underlies comprehension. We aim at understanding how sensory input in general, and speech in particular, is perceived by biological organisms, and at implementing this knowledge in machines. We propose a cortical computation mechanism with cascaded oscillators at the core and test its capability to predict behavioral and neurophysiological performances on challenging acoustic/linguistic conditions. The model provides a metric based on cortical function to assess the interaction between the sensory information stream and the resulting behavior.						
15. SUBJECT TERMS						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	
a. REPORT	b. ABSTRACT	c. THIS PAGE			PATRICK BRADSHAW	
U	U	U	UU	25	19b. TELEPHONE NUMBER (Include area code) 425-8492	

Standard Form 298 (Rev.8/98)
Prescribed by ANSI Std. Z39.18

EXECUTIVE SUMMARY

The objective of our research program is to formulate a fundamental cortical computation principle, universal across sensory modalities, which underlies comprehension. We aim at understanding how sensory input in general, and speech in particular, is perceived by biological organisms, and at implementing this knowledge in machines. Motivated by the possible role of brain rhythms in cortical function, we postulate a cortical computation principle by which the decoding of the sensory information stream is performed within a time-varying window structure, synchronized with the input on multiple time scales. The windows are generated by a segmentation process, implemented by an array of cascaded oscillators. Correct segmentation is a critical prerequisite for correct decoding, and segmentation is correct as long as the oscillators successfully track the input rhythms. Staying in sync with the quasi-regular rhythmicity of the input stream requires that the oscillators be “flexible”, in contrast to autonomous, “rigid” oscillators. We take a reductionist approach and validate the concept within the case study of spoken language comprehension. Previous AFOSR-sponsored study (“Cascading oscillators in decoding speech: A reflection of a cortical computation principle,” Grant number FA9550-11-1-0122) focused on **syllabic** segmentation and introduced a model (TEMPO) capable of explaining a variety of psychophysical and neuroimaging data difficult to explain by current models of speech perception but emerging naturally from the architecture of the model (e.g., Ghitza & Greenberg 2009; Ghitza 2012; Doelling et al. 2014; Ghitza, 2014). The segmentation process in the syllabic level is realized by a flexible theta oscillator locked to the input syllabic rhythm, with the theta cycles constituting syllable-sized windows. The current study focused on the cortical mechanism involved in segmentation at the **phrasal** level. We find that cortical delta oscillations are present only if the input phrase-sized-chunk rate is inside the delta range. This suggests that at the phrase time scale, a delta oscillator plays a role analogous to that of the theta oscillator at the syllable level. The segmentation process is realized by a *flexible delta oscillator* locked to accentuation cues, with delta cycles constituting phrase-sized windows. We claim that, from a functional role perspective, speech decoding is a process paced by an internal, hierarchical clock with *uniform* cortical time units, a theta unit with duration of one Newtonian theta cycle, and a delta unit, with duration of one Newtonian delta cycle. The cortical time units are synchronized with the input. Necessary conditions emerge according to which, in order to stay in lockstep with the input the sequence of operations to decode one syllable *must* be performed within one cortical theta unit, and the sequence of operations to parse one phrase *must* be performed within one cortical delta unit. Importantly, these necessary conditions hold for any decoding computation strategy that may be in place, whether context-invoked or not, whether sequential or not, or whether oscillation driven or not. Hence, the scaffold for the speech decoding process is an acoustic determinant, realized by acoustic-driven theta and delta oscillators. The concept should generalize to all sensory modalities and provides a metric, based on cortical function, to assess the interaction between the sensory information stream and the resulting behavioral actions of the biological organism.

Accomplishments:

(See Section 3 for a detailed description of the research behind the accomplishment listed here.)

- Conceptual
 - Established the role of acoustic-driven theta- and delta- oscillators in decoding speech by a series of behavioral and neurophysiological measurements.
 - Defined the theta and delta cortical units – units of speech information defined by cortical function (Ghitza, 2020).
- Behavioral
 - Provided behavioral evidence for the role of acoustic-driven delta rhythms in setting prosodic markers (Ghitza, 2017).

- Established that the difficulties older adults have in perceiving speech when listening to fast speech is due to a shift of the theta frequency range downwards (Penn et al., 2018).
- Established that neuronal oscillators guide perception of words (Bosker & Ghitza., 2018).
- Neuroimaging
 - Provided neuroimaging (MEG) validation for TEMPO's predictions on the role of acoustic landmarks in driving theta oscillations to facilitate perceptual parsing (Rimmele et al., 2021).
 - Showed that ECoG responses to time-compressed speech vary across the cortical auditory hierarchy (Davidesco et al., 2018).
 - Characterized the dynamics of functional networks for syllable and word-level processing (Rimmele et al., 2021).
- Infrastructure
 - Developed a corpus with systematic coverage of contextually constrained sentences entitled NUBUC (Lewis et al, 2021).
- Analytic Tools
 - Developed XCOV, in MATLAB (Rimmele et al., 2021).
 - Developed computational models for a modulation-spectrum cube, in MATLAB (Unpublished).
- Symposia
 - Organized (with David Poeppel) one-day workshops entitled “Brain Rhythms and Cortical Computation” (BryCoCo). A bi-yearly event.

Publications:

1. Ghitza, O. (2020). "Acoustic-driven oscillators as cortical pacemaker" *Language, Cognition and Neuroscience*. <https://doi.org/10.1080/23273798.2020.1737720>
2. Ghitza O (2017) Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*, 32 (5), 545-561. <https://doi.org/10.1080/23273798.2016.1232419>
3. Penn L, Ayasse ND, Wingfield A, Ghitza O (2018) The possible role of brain rhythms in perceiving fast speech: Evidence from adult aging. *J. Acoust. Soc. Am.* 144 (4), 2088-2094. <https://doi.org/10.1121/1.5054905>
4. Bosker HR, Ghitza O (2018) Entrained theta oscillations guide perception of subsequent speech: behavioral evidence from rate normalization. *Language, Cognition and Neuroscience*, 33 (8), 955-967. <https://doi.org/10.1080/23273798.2018.1439179>
5. Rimmele JM, Poeppel D, Ghitza O (2021). "Acoustically driven cortical delta oscillations underpin prosodic chunking." *eNeuro*. <https://doi.org/10.1523/ENEURO.0562-20.2021>
6. Davidesco I, Thesen T, Honey CJ, Melloni L, Doyle W, Devinsky O, Ghitza O, Schroeder C, Poeppel D, Hasson U (2018) Electrographic responses to time-compressed speech vary across the cortical auditory hierarchy. *bioRxiv*. <https://doi.org/10.1101/354464>
7. Rimmele JM, Sun Y, Michalareas G, Ghitza O, Poeppel D (2019) Dynamics of functional networks for syllable and word-level processing. *Neurobiology of Language* (under review). *bioRxiv*. <https://doi.org/10.1101/584375>
8. Lewis G, van Rijn P, Gwilliams L, Larrouy-Maestri P, Poeppel D, Ghitza O (2021) NYU-BU contextually constrained sentences Corpus: NUBUC. *Zenodo Data Repository*. <https://zenodo.org/record/4075183#.YKVMoi0RrmF>

Personnel:

1. Ghitza, Oded (BU*). PI.
2. Poeppel, David (MPI* & NYU*). Collaboration to publications 1 through 8.
3. Rimmele, Johanna (MPI). Collaboration to publications 5 and 7.
4. Davidesco, Ido (NYU). Collaboration to publication 6.
5. Bosker, Hans (MPI). Collaboration to publication 4.
6. Lewis, Gwyneth (NYU). Collaboration to publication 8.
7. Penn, Lana (Brandeis University). Graduate student to publication 3.

8. Van Rijn, Pol (MPI). Graduate student to publication 8.

*BU = Boston University; *NYU = New-York University; *MPI = Max-Planck Institute

Impact:

1. The computation principle we postulated advances the understanding of the human cortex and is directly relevant for improving man-machine interaction.
2. The concept is validated within the case study of spoken language comprehension but should generalize to all sensory modalities.
3. In terms of our understanding of the human cortex, TEMPO sits at the nexus of an interdisciplinary, coordinated effort (Ghitza, Poeppel, Giraud, Lakatos) to validate the hypothesis that oscillatory properties of cortical circuitry lie at the foundation of many perceptual and attentional phenomena, speech perception in particular. The specific predictions generated by TEMPO facilitate a coherent and constructive back and forth between formal model, stimulus design, psychophysical testing, neurophysiology, imaging – and back.
4. In terms of machine understanding, TEMPO addresses a major weakness common to current, state-of-the-art human language technologies: compared to humans, their accuracy degrades sharply when the input differs from the type of data on which the technology was developed or trained, e.g., when language, dialect, genre and domain are changing, or when environmental conditions are worsening. We believe that this problem cannot be solved by incremental engineering and take a transformational research direction. Our approach is to understand how sensory input in general, and speech in particular, is perceived by biological organisms, and to implement this knowledge in machines. Two Air Force specific arenas in which reliable human language technologies are vital are *Voice command and control in the battlefield* and *Signal intelligence*.
5. The challenge to build systems that enable better 3-D airborne navigation and better control of natural flight remains an important goal of the U.S. Air Force. In particular, there is an ongoing interest to understand sensory and sensorimotor computation principles in biological organisms (e.g., insects, bats, migrating birds), in order to implement them in autonomous systems not yet emulated in engineered flight. The models we develop, universal across sensory modalities, incorporate insights regarding relevant brain mechanisms and provide a new lens on cortical principles of computation relevant to the said Air Force challenge. These models provide metrics based on cortical function to assess the interaction between the sensory information stream and the resulting behavioral actions of the biological organism.

1. Background, Motivation

Our objective is to formulate a fundamental cortical computation principle, universal across sensory modalities, which underlies comprehension. In particular, we aim at understanding how sensory input is perceived by biological organisms, and at implementing this knowledge in machines. We take a reductionist approach and validate the concept within the case study of spoken language comprehension. We pursue the hypothesis that a critical component of successfully decoding speech is the process of **prosodic chunking**, and that this is what largely guides the decoding process on the syllabic and phrasal levels. We hypothesize that prosodic chunking facilitates sensory-linguistic pattern matching via an internal clock-like mechanism that governs the processing time required to link sensory input with stored memory patterns. From a functional role perspective, prosodic chunking is a process paced by an internal, hierarchical clock with *uniform* cortical time units synchronized with the input stream. These units are acoustic driven, with a theta unit that is of one theta cycle long in Newtonian time, and a delta unit, with duration of one Newtonian delta cycle. Hence, the scaffold for the speech decoding process is an acoustic determinant, realized by acoustic-driven theta and delta oscillators. In Section 2 we describe a model of the chunking process with cascaded oscillators at the core, and present behavioral and neurophysiological data that support the model.

1.1 Cortical prosodic chunking

Naturally spoken language is a stream of connected sounds, and although the speech acoustics contain no cues regarding the beginning or end of linguistic units a combination of interleaved cues (e.g., pauses, intonation, syllabic stress) are embedded in the acoustic stream. Information, broadly termed ‘accentuation’ (e.g. intonation, stress, pauses), is used by listeners to indicate boundaries associated with linguistic units (Aubanel et al., 2016; Oganian and Chang, 2019). The process by which the input signal is partitioned into temporal segments to be linked to a variety of linguistic levels of abstraction (ranging from phonetic segments to syllables to words and, ultimately, prosodic and intonational phrases) is called ‘segmentation’.

The segmentation process has been shown to operate on time intervals associated with syllables (up to about 250 ms) (Brungart et al., 2007; Doelling et al., 2014; Ghitza and Greenberg, 2009; Kösem et al., 2018), and a similar process has been suggested to operate on the phrasal level (0.5–2 s) (Ding et al., 2016; Ghitza, 2017; Keitel et al., 2018; Martin and Doumas, 2017). At the syllabic level, perceptual segmentation—or chunking—is by and large a pre-lexical process. Oscillation-based models propose that this segmentation is realized by flexible theta oscillations aligning their phase to the input syllabic rhythm (‘speech tracking’), where the theta cycles mark the speech chunks to be decoded (e.g., Ahissar and Ahissar, 2005; Assaneo et al., 2020; Ding and Simon, 2009; Ghitza, 2011; Giraud and Poeppel, 2012; Gross et al., 2013; Haegens and Zion Golumbic, 2017; Hovsepian et al., 2020; Lakatos et al., 2019, 2005; Peelle and Davis, 2012; Pittman-Polletta et al., 2020; Poeppel, 2003; Rimmele et al., 2018). At the phrase level, phrase rhythm can affect segmentation (Gee and Grosjean, 1983; Martin, 2015; Deniz and Fodor, 2019; Hilton and Goldwater, 2020). There have been various studies aiming to quantify phrase length and rhythmicity (e.g., Breen, 2018; Clifton et al., 2006; Deniz and Fodor, 2019), suggesting that typical intonational phrases are about one second in duration (Auer et al., 1999; Inbar et al., 2020; Stehwien and Meyer, 2021). More specifically, the duration of intonational phrases spans a range between approximately 0.5 to 1.5 sec in English (slightly faster in some other languages) (Inbar et al., 2020; Stehwien and Meyer, 2021). Prosodic segmentation (here also termed ‘prosodic chunking’) is based on intonation units that contain specific prosodic cues (such as pauses or pitch contour), which can pace the information flow at the phrasal time scale (Inbar et al., 2020; Shattuck-Hufnagel and Turk, 1996). The extent to which phrase level rhythmic structure supports segmentation and structural parsing was not widely studied. Here, we investigate the neural processing of rhythmic structure at

the phrasal scale by analyzing how individuals' group single digits into 'phrase-sized' digit chunks. What kind of neuronal mechanism can realize this chunking process?

Cortical delta oscillators – with a frequency range ($\sim 0.5\text{--}2$ Hz) that corresponds to the phrasal time scale – were shown to be elicited during phrasal processing of speech or chunking processes at the phrasal scale (Buiatti et al., 2009; Ding et al., 2016; Meyer et al., 2016; Bonhage et al., 2017; Keitel et al., 2018; Boucher et al., 2019). Delta was observed in the posterior superior temporal, the inferior frontal, precentral, and temporal-parietal cortex using ECoG (Ding et al., 2016), and using EEG at bilateral middle and superior temporal areas (also fusiform gyrus) (Bonhage et al., 2017) and at fronto-temporal sites (Boucher et al., 2019). (Recall the ambiguous definition of the delta range in the literature, which covers a range of overlapping frequency bands inside the 0.5 to 4 Hz frequency range (e.g., Bonhage et al., 2017; Bröhl and Kayser, 2020; Keitel et al., 2018). Since we are interested in the segmentation of phrasal chunks, which in English are about 0.5 to 1.5 sec long (e.g., Miller, 1962; Inbar et al., 2020), we opted to define the delta frequency band to be 0.5–2 Hz).

1.2 Speech technology

Many speech technologies are already useful and constantly improving, but they share a common weakness: compared to humans, their accuracy degrades sharply when the input differs from the type of data on which the technology was developed or trained. Most state-of-the-art language understanding systems employ modeling and pattern recognition techniques which require massive amounts of speech data, along with transcriptions, annotations, analyses, parses, or other metadata to train the internal states. In comparing the computation principles utilized by current automatic speech recognition (ASR), and language understanding systems, to the current knowledge we have on how sensory input in general (and speech, in particular) is perceived by biological organisms, remarkable differences are noted:

Theta chunking and ASR: Current ASR systems operate in uniform time intervals: acoustic feature vectors are computed once approximately every 20 ms; and sub-word units, each represented as an ordered sequence of those feature vectors, are recognized using a statistical pattern recognition framework. *In cortical computation, on the other hand, the decoding process is executed inside 'temporal windows,' where a window is one cycle of a theta oscillator locked to the input rhythm. Considering such acoustic-driven window as a 'cortical-time unit (CTU)', cortical recognition operates in uniform cortical-time units: a theta-syllable object (a VCV) is computed once every one theta CTU (Ghitza, 2013).* In part inspired by this cortical computation principle, developed in our recent work, Räsänen et al. (2015) presented a syllable-based approach to unsupervised word discovery from speech. An oscillator-based algorithm was implemented and used for unsupervised syllabic segmentation (i.e., theta parsing). Feasibility of the approach was investigated on spontaneous American English and Tsonga language samples, with promising results.

Delta chunking and language understanding systems: In current language understanding systems, rules of parsing are derived from linguistic structure acquired by hand labeling of phrasal units over large corpora, e.g., in the form of N-gram statistics (e.g., Levit et al. 2002; Levit 2004). *In cortical computation, phrase constituent candidates may be computed within temporal windows that are one cycle of a delta oscillator locked to the phrasal structure of the input signal (a 'delta CTU'): as the result of the evolutionary trajectory to match a cortical function, the externalization of linguistic structure evolved to match constraints imposed by the delta chunking circuitry in the cortical receiver layers, in analogy to the evolution of the syllabic structure in an attempt to match constraints imposed by the theta parsing circuitry in the pre-lexical layers.*

1.3 Collaborative effort

The hypothesis that oscillatory properties of cortical circuitry may lie at the foundation of many perceptual and attentional phenomena also motivates the research programs of David Poeppel at

Max-Planck Institute (MPI) and NYU. Each approach championed by our two labs (psychophysics, neurophysiology, computational modeling) has proceeded independently and productively, but we have established that a joint and shared attack will move us significantly forward in a synergistic manner, in particular because it forces engagement with the issues across methods and across levels of analysis. The key conceptual link that we are advancing—that yields new insights well beyond what the results of each lab can contribute individually—is provided by TEMPO (see review in subsection 2.1), the increasingly detailed computational model that now sits at the nexus of the research program. The specific predictions generated by TEMPO facilitates a coherent and constructive back and forth between formal model, stimulus design, psychophysical testing, neurophysiology, imaging – and back. The interdisciplinary approach we advocate is facilitated by the adoption of a model that captures a foundational principle about cortical computation, accounts for complex behavioral data, is supported by MEG data, and that can be clearly articulated and tested across multiple levels.

In the past eight years we made substantial progress on two issues. It is the central conjecture motivating this proposal that (a) an oscillation-based framework depicted in Figs. 1 and 2 provides an innovative approach to the cognitive, computational, and neural challenges of how sensory input in general—and speech in particular—is perceived by biological organisms, and therefore merits detailed empirical characterization and (b) that such a model provides an important algorithmic and implementational set of constraints to understand the role of prosodic chunking in the interaction between the sensory information stream and the resulting behavior.

2. Segmentation with nested oscillations

Ghitza, O. (2020). "Acoustic-driven oscillators as cortical pacemaker" *Language, Cognition and Neuroscience*. <https://doi.org/10.1080/23273798.2020.1737720>

2.1 Role of oscillators — current view

Speech (*everyday* speech, in particular) is inherently a quasi-rhythmic phenomenon in which the talker's linguistic information is transmitted in “packets”, manifested in the acoustic signal in the form of temporal “chunks”. Oscillation-based models of the speech process postulate a cortical computation principle by which, the *decoding* process is performed on acoustic chunks defined by a time-varying window structure synchronized with the input on multiple time scales. In the following we shall exemplify this computation principle with TEMPO (Ghitza, 2011), a model which epitomizes recently proposed oscillation-based models of speech perception (e.g., Poeppel 2003; Ahissar & Ahissar 2005; Ding & Simon 2009; Ghitza & Greenberg 2009; Giraud & Poeppel 2012; Gross et al., 2013; Peelle & Davis 2012).

The model is shown in Figure 1. The sensory stream (generated by a model of the auditory periphery, e.g., Chi et al., 1999; Messing et al., 2009) is processed, simultaneously, by a *segmentation* path and a *decoding* path. The segmentation path is in blue. The theta and delta oscillators are flexible, e.g., the VCO component in the classical PLL circuit (Viterbi, 1966; Ahissar et al., 1997; see also the biophysical computational model by Pittman-Polleta et al., 2020), with quasi-periodic oscillations that are locked to the quasi-rhythmic acoustic syllable and phrasal chunks.

The decoding path is in orange. Decoding is steered by segmentation: the decoding process evolves within the theta/delta cycles. Conventional models of speech perception assume a strict decoding of the acoustic signal¹. The decoding path of TEMPO, which links acoustic chunks of different durations with stored linguistic memory patterns, conforms to this notion. Not present in

¹ In conventional models of speech perception phones are identified first, and the ordered sequence of identified phonemes results in a pointer to the word lexicon (e.g., Marslen-Wilson, 1987; Luce & McLennan, 2005; Stevens, 2005).

conventional models is the segmentation path, which determines the acoustic chunks (their location and duration) to be decoded. As it turns out, segmentation plays a crucial role in explaining a range of counterintuitive psychophysical data that are hard to explain by the conventional models (e.g., Ghitza & Greenberg, 2009; Ghitza, 2012, 2014 & 2017). In TEMPO, the segmentation path is realized by an array of flexible oscillators locked to the input rhythm.

In the pre-lexical level of TEMPO, the segmentation process is realized by a *flexible theta oscillator* locked to the input syllabic rhythm, where the theta cycles constitute the syllabic windows. A theta cycle is set by an evolving phase-locking process (e.g., a PLL circuit, Viterbi, 1966; Ahissar et al., 1997), during which the code is generated. Doelling et al. (2014) provided MEG evidence for the role of theta, showing that intelligibility is correlated with the existence of acoustic-driven theta neuronal oscillations.

In the phrase level, the segmentation process is realized by a *flexible delta oscillator* locked to the input phrase-chunk rhythm, where the delta cycles constitute the phrase-chunk windows. A delta cycle is set by an evolving phase-locking process, during which contextual parsing proceed. Rimmele et al. (2021) provided MEG evidence for the role of acoustic-driven delta, showing that the accuracy of digit retrieval is correlated with the existence of acoustic-driven delta neuronal oscillations.

2.2 Role of oscillators — a broader look

As seen in subsection 2.1, the functional role of the acoustic-driven theta and delta oscillators is to facilitate a time-varying window structure, synchronized with the input, where the theta/delta cycles determine the syllable/phrase chunks to be decoded. In this subsection, a broader functional role for the acoustic-driven theta and delta is postulated, namely, they constitute an internal, hierarchical clock that pace the speech decoding process to stay in sync with the linguistic information flow, via keeping the decoding process operating on acoustic chunks aligned with proper linguistic units. In the following I shall outline the rationale for this postulate.

In Figure 2a, the sequence of operations that are executed in mapping the acoustic stream onto a series of syllable objects is outlined in more detail. First is the segmentation process, in the form of acoustic-driven theta cycle, set by an evolving phase-locking process² (step 1 in Fig. 2a). While the theta cycle is evolving, a neural code for the syllable chunk is generated throughout the theta cycle, e.g., in the form of gamma nested in theta³ (step 2). The code is transmitted at the end of the theta cycle (step 3), then recognized (i.e., a working memory storage is activated) during the next theta cycle (step 4). An additional functional role of theta—beyond the setting of the theta window—emerges: the end-time of the theta cycle marks the moment at which the code is transmitted, i.e., it marks the moment by which the code generation must end. This is a necessary condition because, beyond this moment, the code-generation circuitry should already be occupied with the generation of the code for the next theta chunk.

Turning to the phrase level, the sequence of operations that take place in mapping the stream of syllable objects onto phrase constituent candidates (PCCs) is shown in Fig. 2b. Segmentation comes first, in the form of acoustic-driven delta set by an evolving phase-locking process⁴ (step 1 in Fig. 2b). While the delta cycle is evolving, PCCs are obtained by a parsing process that take place throughout the delta cycle (step 2). The PCCs are multiplexed at the end of the delta cycle (step 3). An additional functional role of delta emerges, analogous to that of theta: the end-time of the delta

² The acoustic cues to which the theta oscillator is locked to are still under debate (acoustic edges? vocalic nuclei?).

Here, the theta cycle is locked to vocalic nuclei, hence the syllable objects are in the form of VCVs (Ghitza, 2013).

³ A possible mechanism to generate the neural code is via gamma sampling (Shamir et al., 2009; Ghitza, 2011).

⁴ The delta oscillator is locked to *accentuation* attributes; the acoustic cues that form accentuation are still under pursuit.

cycle marks the moment by which the PCCs must be delivered. Three points merit discussion. First, while we find cortical oscillations with cycle durations that correspond to syllables and phrases (theta and delta), we don't have oscillations that correspond to words. Indeed, there is no compelling linguistic evidence that words are regular enough for phase locking. Therefore, in TEMPO, the lexical access process operates on the syllable stream without any segmentation-based supervision (see, for example, the model TRACE, Luce & McLennan, 2005). Second, in generating the PCCs, numerous computation strategies can be considered (e.g., template matching; statistical pattern recognition; predictive coding; inference Bayesian approach; analysis-by-synthesis; relations via correlations; statistical learning). The parsing process operates on words throughout the delta window and is not necessarily sequential, nor it is necessarily oscillation based⁵. Important to our discussion, regardless of the computation strategy, in order to stay in lockstep with the input information flow the derivation of the PCCs must be concluded by the end of the delta window. And third, in deriving language features, including syntax and sentence-level semantics, the language process operates on a sequence of PCCs that span a few delta cycles. A few questions—beyond the scope of this commentary—remain open, e.g.: how the duration of the language process “window” is determined? Is it formed by a segmentation process realized by ultra-slow oscillators locked to the sentence-level information flow?

2.3 Cortical time units (CTUs)

The sequence of operations described in subsection 2.2, in the syllable level and in the phrase level, is repetitive, irrespective of the theta/delta window durations. Functionally, therefore, the speech decoding process can be viewed as a process paced by an internal clock with uniform cortical time units (CTUs): (i) a “theta unit”, with duration—in Newtonian time⁶—of one theta cycle, and (ii) a “delta unit”, with duration of one delta cycle. The CTUs are set by oscillators that are in sync with the input. As such, the CTUs, uniform in the internal domain, span non-uniform durations in Newtonian time (Fig. 2c, blue boxes). Crucially, the CTUs have a limited range, bounded in Newtonian time by the upper frequency range of the oscillators. Hence, the shortest duration of a theta CTU is about 125 ms (for $\theta_{\max} = 8$ Hz), and the shortest duration of a delta CTU is about 0.5 sec (for $\delta_{\max} = 2$ Hz). Speech decoding, therefore, is viewed as a process that proceeds in uniform cortical-time ticks: at the syllable level, the entire sequence of operations in Fig. 2a is executed in one theta CTU; at the phrase level, the entire sequence of operations in Fig. 2b is executed in one delta CTU. (Newtonian time and cortical time, at the syllable level, are illustrated in Fig. 2c.)

3. Accomplishment during past 40 months

3.1 Behavioral

3.1.1 Acoustic-driven delta rhythms as prosodic markers.

Ghitza O (2017) Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*, 32 (5), 545-561. <https://doi.org/10.1080/23273798.2016.1232419>

In this study we asked whether the cortical computation principle epitomized in the syllable level—that a (syllabic) segmentation process guides the syllable decoding process—can be generalized to the chunk level, where a chunk is a speech fragment that is multi-word in duration. We hypothesize that acoustic prosodic segmentation guides the phrase decoding process, using a flexible delta oscillator playing a role analogous to the flexible theta oscillator in syllabic segmentation.

To test this hypothesis we conducted two psychophysical experiments with four questions in mind: (1) Does acoustic prosodic segmentation play a role in word retrieval? (2) Does the chunking

⁵ The role of endogenous oscillations in generating abstract linguistic predictions (e.g., Meyer & Gumbert, 2018) is still under debate.

⁶ Newtonian time, in seconds. See Chapter “Newtonian and Bergsonian Time,” in Wiener, 1948.

strategy---how elements are grouped into chunks---play a role in digit retrieval?⁷ (3) Are hidden prosody cues---e.g., accentuations arching over a chunk---as effective as explicit temporal grouping (by gap insertions in between the chunks)? (4) Is acoustic prosodic segmentation driven by a delta oscillator? In both experiments, listeners heard 10-digit utterances with different chunking patterns and prosody modes, followed by 2- and 3-digit-long targets, and were asked to indicate whether the target was part of the preceded utterance (i.e., subjects performed a retrieval task in the form of a target identification task).

Experiment 1 addressed the first three questions. Error rate was measured as a function of *chunking pattern* and *prosodic mode*. Chunking rate was inside the cortical delta frequency range for all chunking patterns considered, eliminating the chunking rate as a factor. Fifteen chunking patterns, listed in the abscissa of Fig. 3b, were used. The chunking patterns form regular and irregular grouping patterns, e.g., 22222 (regular) or 424 (irregular)⁸. In generating the acoustic waveforms two prosody modes were used for all chunking patterns, a *Gapped* mode and an *Accentuated* mode. Fig. 3a shows the resulting stimuli (waveform and a Fourier spectrogram) for root string 3323443215 synthesized in the Gapped mode (top), and in the Accentuated mode (bottom), with the chunking pattern 3322. Note that a stimulus synthesized in the Gapped mode belongs to the 'temporal grouping' condition of Ryan (1969), and a stimulus synthesized in the Accentuated mode could be considered as belonging to Ryan's (1969) 'non-temporal grouping' condition since no explicit grouping have been applied. However, an implicit grouping exists, enforced by accentuation.

Participants performed a digit retrieval task, in the form of an adapted Sternberg target identification task (Sternberg, 1966) (target ID task from here on): they heard a 10-digit stimulus followed by a 2- or a 3-digit long target, and were asked to indicate whether or not the target was part of the utterance. Error rate was measured for three target positions: *Target Inside* a chunk, *Target Split* between two successive chunks, and *No Target* present in the 10-digit string.

Experiment 1, results. Irrespective of the chunking pattern, performance is high only for targets inside of a chunk, pointing to the benefit of acoustic prosodic segmentation in digit retrieval. Fig. 3b show the error rates for Target Inside. Similar error patterns emerge for all chunking patterns considered, indicating that as long as the chunking rate is inside the cortical delta range, chunking strategy is not an important factor. Similar error patterns also emerge for the two prosody modes used for chunking, indicating that (natural) accentuation is as effective as gap insertions in enabling effective segmentation.

Experiment 2 addressed the fourth question. The role of cortical delta was tested by driving chunking rate from inside to outside of the cortical delta range. The experimental paradigm was the same as in Experiment I. The 10-digit utterances were chunked with prescribed chunking rates, and performance was measured as a function of the rate. Stimuli were generated for a number of conditions specified by two parameters, *chunking pattern* and *chunking rate*. Four chunking patterns were used: the regular patterns 22222 and 333, and the irregular patterns 3322 and 2233 grouping. Chunking rate was controlled by varying the silent gaps inserted in between the chunks. Eight gap durations were used: 1, 20, 40, 80, 120, 160, 180, 200 ms, providing a gradual change in chunking rate -- in and out of the delta frequency band⁹. Fig. 3c shows the resulting stimuli for the root string 1058658263 synthesized in Accentuated mode and chunking pattern 22222, with a 20-

⁷ For example, why do Americans chunk a 10-digit telephone number into groups of 3322? and why, in Europe, some use a chunking pattern of 22222?

⁸ To illustrate our notation of chunking pattern, the root string 3762895069, for example, can be chunked into the regular chunking pattern 22222 [37 62 89 50 69], or into the irregular pattern 3322 [376 289 50 69], etc.

⁹ To be able to generate stimuli with chunking rates greater than the upper bound of the delta frequency range (about 3 Hz), chunk (and target) waveforms were time compressed by a factor of 3 -- just below auditory channel capacity (Ghitza, 2014).

ms long silent gap (top) and a 160-ms long gap (bottom) and corresponding chunking rates of 4.2 Hz (top) and 2.8 Hz (bottom). Note that the waveform of each chunk is time-compressed by a factor of 3. Finally, error rate was measured only for the Target Inside condition.

Experiment 2, results. As shown in Fig. 3d, performance remains high as long as the chunking rate is inside the frequency range of neuronal delta and it sharply deteriorates once the chunking rate is higher than the upper limit of delta, indicating a possible role of delta in acoustic prosodic parsing.

3.1.2 The possible role of brain rhythms in perceiving fast speech: Evidence from adult aging.

Penn L, Ayasse ND, Wingfield A, Ghitza O (2018) The possible role of brain rhythms in perceiving fast speech: Evidence from adult aging. *J. Acoust. Soc. Am.* 144 (4), 2088-2094.

<https://doi.org/10.1121/1.5054905>

Previous research has demonstrated that in young adults, the intelligibility of time-compressed speech can be rescued by ‘repackaging’ the speech signal through the regular insertion of silent gaps to restore correspondence to the theta oscillator (Ghitza and Greenberg, 2009). Compared to young adults, intelligibility by older adults is affected much more by an increase in speech speed (e.g., Gordon-Salant and Fitzgibbins, 1993; Wingfield et al., 1999). One might raise the question of whether the “repackaging” effect can be extended to populations that have different neuronal properties from the typical young adult. It has been suggested, for example, that aging is associated with changes in slow wave neural oscillations (e.g., delta and theta waves). If older adults exhibit an earlier and steeper decline in intelligibility, and an earlier “rescue” effect of repackaging, it would be because of slower oscillations, theta in particular. If such results were found, the model TEMPO (Ghitza, 2011) would see this in terms of a shift in the oscillator’s frequency range, such that cortical theta oscillations are slower in older compared to younger adults.

This study addressed this question by conducting two experiments using spoken random-digit strings. Digits strings were used in order to eliminate the effect of linguistic context and to focus exclusively on bottom-up processes. In Experiment 1, the accuracy of digit recognition as a function of speech speed by young and older adults was recorded and served as baseline performance. The speech rates used ranged from 6 to 21 syllables per second. In agreement with other studies, we would expect accuracy to drop with increasing speech rates and with a steeper rate of decline for the older adults. Of special interest, however, is the extent to which repackaging of the speeded speech results in intelligibility recovery. In Experiment 2, time-compressed digit strings, as described above, were repackaged, with packaging rates ranging from 2 to 7 packets per second (corresponding to the theta frequency range). Following the TEMPO model, we postulated that repackaging would rescue intelligibility for both young and older adults, with the recovery for the older adults appearing at a lower packaging rate.

Experiment 1, stimuli. The experimental stimuli consisted of 100 spoken digit strings of 4 digits recorded by a male speaker of American English at a rate of 3 syllables per second. The four items in each digit list were randomly drawn from the digits 0 to 9. Each list was preceded by four, 400 Hz tones, at a rate designed to provide an opportunity for the presumed cortical theta oscillator to entrain to the input rhythm of the digit stimuli prior to their presentation.

The waveforms of the recorded digit strings were time-compressed using a pitch-synchronous, overlap and add (PSOLA) procedure (Moulines and Charpentier, 1990) incorporated into Praat (Boersma & Weenink, 2016), a speech analysis and modification package¹⁰. The compression factor was gradually increased to generate time-compressed versions the digit strings, labeled by speech speeds measured in terms of syllables per second. Six speech speeds were used in this study:

¹⁰ With this method the formant patterns and other spectral properties of the time-compressed signal are preserved but altered in duration. The fundamental frequency (“pitch”) contour, however, remains the same.

6, 9, 12, 15, 18, and 21 syllables per second. Six versions of the core stimuli (sets of 4-digit lists preceded by four entrainment tones) were created, one for each speech rate, by concatenating the time-compressed digit strings with four tone packets at a packet frequency that is equal to the corresponding speech speed. For example, a digit string with a speed of 12 syllables per second would be preceded by four tone packets at 12 Hz packet frequency, recorded with a 50% duty cycle (about a 40 ms packet duration) and a 400 Hz tone inside the packet (see Fig. 4a, top).

Experiment 2, stimuli. “Repackaging” refers to the process of dividing the time-compressed waveform into fragments, called packets, and delivering the packets at a prescribed rate (Ghitza and Greenberg, 2009). Figure 4a (middle) illustrates the repackaging process. The upper panel shows the waveform of a sentence time-compressed by a factor of $\kappa = 3$. The compressed waveform is blindly segmented into packets with equal duration of δ (grey boxes). The lower panel shows the time-compressed waveform after repackaging, with a packaging rate of $1/T$ packets per second (or Hz). The acoustic signal inside the δ -long packet is the time-compressed signal. An exhaustive study on the interaction between the parameters of the repackaging process and its effect on the intelligibility and the quality of the speech output is yet to be conducted (see a few special cases in Ghitza & Greenberg, 2009; Ghitza, 2014). Although this would be an important area for future research, here we used a setting in which the duty cycle (δ/T in Fig. 4a, middle) is constant.

Six repackaging conditions were used in this study corresponding to the six speech speeds determined by the uniform time compression: 2, 3, 4, 5, 6, and 7 packets per second, with the speech rate inside a packet being 6, 9, 12, 15, 18, and 21 syllables per second, respectively. The packet duration was set to result in a constant duty cycle of 20% across the six packaging rates. Fig. 4a (bottom) shows example waveforms of repackaged digit-list stimuli at 2, 4, and 7 packets per second, with packets containing 6, 12, and 21 syllables per second uniformly time-compressed, respectively, each preceded by a sequence of tone packets with a packet frequency corresponding to the packaging rate (see Fig. 4a, bottom).

Experiments 1 and 2, procedure. Each participant received 120 experimental trials, each composed of a 4-digit string preceded by 4 tones. Half of the digit strings were time-compressed without repackaging while the other half were heard with the same compressed syllabic rate but with a repackaging rate at the respective κ . In order to keep the study at a manageable length and to avoid fatigue, half of the young and older adult participants received rates of 6, 12, and 18 syllables per second while the other half received rates of 9, 15, and 21 syllables per second for both the compressed and repackaged conditions. Each condition was preceded by 3 practice trials.

The repackaged and time-compressed conditions were presented in a blocked design, with the order of conditions counterbalanced across participants. The order of presentation of speech rates within conditions was also blocked, with the order of rates varied between participants.

Experiments 1 and 2, results. Figure 4b shows the percentage of 4-item digit-lists reported correctly and in the correct order by the young and older adults as time-compression was used to systematically increase speech rate (indicated on the lower abscissa), and when the compressed signals were repackaged, with the repackaging rate, in packets per second (indicated in the upper abscissa). Note that the syllable rate indicated in parentheses in the upper abscissa is the same as the rates indicated in the lower abscissa.

Our data show that, in the baseline time-compression condition, performance by the older adults was characterized by a steep decline in accuracy as a function of speech rate, a finding consistent with previous studies. In the repackaging condition it was found that recovery occurs only at speeds higher than a “rescue point,” and performance at that region remains steady (i.e., the size of intelligibility recovery grows as waveform speed increases). This behavior is in line with the notion that, when listening to time-compressed speech, additional decoding time is needed (e.g., Ghitza and Greenberg, 2009; Vagharchakian et al., 2012), determined by neuronal entrainment to theta. For the young adults, the rescue point occurs at 5 packets per second – at the center of the theta frequency range for typical young adults. This is in line with the assertion that the optimal range of packaging rate is inside the range of theta, and that the best synchronization between the

speech stream and theta is achieved by tuning the packaging rate towards the mid-range of theta (Ghitza, 2011). Importantly, by contrast with the young adults, the rescue point for the older adults occurred by 4 packets per second, distinctively lower than that of the young adults. Interpreted through the prism of TEMPO, this result implies a decrease in the upper limit of the frequency range of the oscillations, or perhaps a potential shift downward in the frequency range of the theta oscillations with adult aging. In sum, this study provides behavioral findings that are consistent with a postulated role of neuronal oscillations in speech perception, exemplified by the TEMPO model.

3.1.3 Entrained theta oscillations guide perception of subsequent speech.

Bosker HR, Ghitza O (2018) Entrained theta oscillations guide perception of subsequent speech: behavioral evidence from rate normalization. *Language, Cognition and Neuroscience*, 33 (8), 955-967.

<https://doi.org/10.1080/23273798.2018.1439179>

In this study we tested whether neural entrainment is at the core of the ‘rate normalization’ process. When a listener is required to identify a target speech sound ambiguous between two members of a durational contrast (e.g., in English, short /b/ vs. long /w/; in Dutch, short /a/ vs. long /a:/), his decision may be biased towards the longer phoneme (i.e., /w/ in English; /a:/ in Dutch) if presented after a preceding sentence (hereafter: carrier) produced at a faster speech rate (Kidd, 1989; Pickett & Decker, 1960; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). In this study we hypothesize that the rate normalization process is shaped only by carrier sentences capable to elicit neural entrainment in the theta range (3-9 Hz).

To test this hypothesis we conducted three psychophysical experiments. In Experiment 1, the finding that auditory channel capacity is about 9 syllables/sec, determined by the upper frequency limit of theta (Ghitza, 2014) was replicated with Dutch materials. In Experiment 2, Dutch participants were presented with time-compressed carriers followed by uncompressed target words containing vowels ambiguous between Dutch /a/ and /a:/. The carriers were linearly compressed by various compression factors κ , resulting in some syllable rates within theta range (i.e., below 9 Hz) and some outside theta range (i.e., above 9 Hz). In this experiment we test whether the effectiveness of the rate normalization process deteriorates for carriers with syllabic rate above 9 Hz. As we shall see, indeed, Experiment 2 show that rate normalization effects are only triggered by syllable rates within theta range. In Experiment 3, we tested whether the effectiveness of rate normalization is restored through repackaging (Ghitza & Greenberg, 2009; Ghitza, 2014). For instance, compressed-speech carrier with a syllable rate of 15 Hz would not be predicted to bias the perception of subsequent Dutch target vowels towards /a:/, because the syllable rate is outside theta. Are rate normalization effects should be restored if this compressed speech carrier would be repackaged such that the packet delivery rate would fall below 9 Hz?

Experiment 1 aimed at replicating the findings from Ghitza (2014) in Dutch, testing the *intelligibility* of Dutch speech recordings compressed with various compression factors κ such that resulting syllable rates would fall either within (i.e., below 9 Hz) or outside theta range (above 9 Hz). Moreover, it tested whether repackaging speech such that the resulting packet delivery rate falls within theta would restore intelligibility of highly compressed speech, as reported in Ghitza (2014). If Experiment 1, targeting overall intelligibility, succeeds in replicating the findings from Ghitza (2014), this allows us to use the Dutch speech materials for the subsequent experiments, targeting rate normalization.

Participants. Native Dutch participants ($N = 21$; 17 females, 4 males; $M_{\text{age}} = 22$) with normal hearing were recruited from the Max Planck Institute’s participant pool. (They gave informed consent as approved by the Ethics Committee.)

Design and materials. Sentence materials were generated after Ghitza (2014). A male native speaker of Dutch was recorded, fluently producing 100 digit-strings. Each digit string was comprised of 7 digits (e.g., *twee een vijf, vier zes vijf drie*; “two one five, four six five three”),

approximately 2 seconds long, with an average syllable rate around 3 Hz ($SD = 0.2$). The speaker uttered the strings as a phone number, using a cluster of 3 digits followed by a cluster of 4 digits. Digit strings were time-compressed using PSOLA in Praat (Boersma & Weenink, 2016). The time-compression factor κ was chosen to be 1, 2 and 5. (Fig. 5a) Signals with $\kappa = 1$ and signals with $\kappa = 2$ had an average syllable rate of around 3 Hz and 6 Hz, respectively. As such, their syllable rates fell below 9 Hz, generally assumed to be the upper limit of cortical theta and the maximum reliable information transfer rate through the auditory channel (i.e., auditory channel capacity; Ghitza, 2014). Signals with $\kappa = 5$ had an average syllable rate of around 15 Hz (i.e., above 9 Hz), falling outside cortical theta and auditory channel capacity. A fourth repackaged condition involved repackaging the $\kappa = 5$ condition (Fig. 5a). Intervals of 66 ms were excised from the signal with $\kappa = 5$ and spaced apart by 100 ms using Praat. This manipulation resulted in a repackaged condition that contained acoustic-phonetic material that was identical to the $\kappa = 5$ condition, only with a significantly lower packet delivery rate of 6 Hz (i.e., within the cortical theta range).

Experiment 1, results. As is shown in Fig. 5b, the data replicates the findings from Ghitza (2014) for Dutch materials. Speech intelligibility was high for syllable rates within the theta range (i.e., $\kappa = 1$ and $\kappa = 2$ with syllable rates below 9 Hz) but deteriorated sharply for syllable rates outside the theta range (i.e., $\kappa = 5$ with syllable rates above 9 Hz). Moreover, when maximally compressed speech (i.e., $\kappa = 5$) was repackaged such that the resulting packet delivery rate falls within theta range, the intelligibility of the speech was restored to a large degree. Thus, Experiment 1 validated the use of these materials for subsequent experiments.

Experiment 2 was designed to test whether there is an upper limit to the carrier syllable rates that may induce rate normalization. Instead of a gradual increase in rate normalization effects as the compression factor (hence the syllable rate) is raised, oscillations-based models predict that rate normalization effects should not be induced by syllable rates above 9 Hz, since cortical theta would then be assumed to be out of sync with the syllabic input rate. Experiment 3 was designed to test whether repackaged compressed speech may induce rate normalization. Since repackaging of heavily compressed speech can lower the packet delivery rate to below 9 Hz (i.e., within theta range), oscillations-based models predict that repackaging might restore rate normalization effects. Here we report on the key results from Experiments 2 and 3, combined. See publication for the full scope of these experiments.

Design and materials. A set of 10 digit-strings (out of the total set of a hundred) were adopted from Experiment 1, to serve as carriers in Experiments 2 and 3. Four carriers were used, three with uniform time compression (with $\kappa = 1, 2, 5$), and one subjected to repackaging (with $\kappa = 5$). Recall that the carrier with $\kappa = 5$ had a syllable rate above 9 Hz, falling outside cortical theta and auditory channel capacity, while the carriers with $\kappa = 1, 2$ and the repackaged carrier fall below the upper limit of cortical theta. These digit-string carriers preceded the target word. These Dutch target words were proper names selected from four different minimal pairs containing either the short vowel /a/ or the long vowel /a:/: *Ad - Aad*, /at - a:t/; *Bas - Baas*, /bas - ba:s/; *Dan - Daan*, /dan - da:n/; *Mart - Maart*, /mart - ma:rt/. The /a/-/a:/ vowel contrast in Dutch is cued by both spectral (lower formant values for /a/, higher formant values for /a:/) and temporal (shorter duration for /a/, longer duration for /a:/; Escudero, Benders, & Lipski, 2009).

Experiments 2 and 3, results. The data is shown in Fig. 5c, suggesting that only when theta oscillations can track the rate of the carrier ($\kappa = 2$) do we find effects on the perception of subsequent target words. When the carrier is compressed to have syllable rates outside the theta range, no difference in target word categorization is observed (Experiment 2). The novel finding of Experiment 3 is that when the $\kappa = 5$ carrier was repackaged such that its packet delivery rate was set to be around 6 Hz (inside theta range), rate normalization was restored. Thus, Experiment 3, together with the replicated findings from Experiment 2, supports an oscillations-based account of rate normalization, with a central role for sustained entrainment of oscillations in the theta range in rate normalization.

In sum, the data show that uniform compression of a speech carrier to syllable rates *inside* the theta range influences perception of subsequent uncompressed targets, but compression *outside* theta range does not. However, the influence of carriers – compressed outside theta range – on target perception is salvaged when carriers are ‘repackaged’ to have a packet rate inside theta. This suggests that the brain can only successfully entrain to syllable/packet rates within theta range, with a causal influence on the perception of subsequent speech, in line with recent neuroimaging data (Doelling et al., 2014; Kösem et al., 2017). Thus, this study points to a central role for sustained theta entrainment in rate normalization and contributes to our understanding of the functional role of brain oscillations in speech perception.

3.2 Neuroimaging

3.2.1 Acoustically driven cortical delta oscillations underpin prosodic chunking.

Rimmele JM, Poeppel D, Ghitza O (2021). "Acoustically driven cortical delta oscillations underpin prosodic chunking." *eNeuro*. <https://doi.org/10.1523/ENEURO.0562-20.2021>

In this study we tested the hypothesis that the decoding process in the phrasal level is guided by a delta oscillator locked to the accentuation acoustic cues, by recording Magnetoencephalography (MEG) data while participants performed a digit retrieval task. This hypothesis was first introduced in the behavioral study by Ghitza (2017; see subsection 3.1.1), and here we aimed at providing the MEG, “smoking gun” validation. The digits in the string were grouped into chunks, with chunk rates either inside or outside to the delta frequency range. The experiment addressed two questions: (1) Do elicited delta brain waves correlate with behavior, such that impaired performance in digit retrieval occurs if the chunk rate is outside of the delta range? (2) Where in the auditory pathway do those neuronal oscillations occur?

Digit string stimuli. We used 10-digit long stimuli where we manipulated the pauses in-between digits according to the experimental conditions. The digit sequences were grouped into chunks, with a chunk pattern termed 2222. For example, the 2222 pattern of the sequence 4259522560 is [42 59 52 25 60]. We used two chunk rates: 1.8 Hz (inside the delta range) and 2.6 Hz (at the outside border of the delta range, referred to as “outside”), termed conditions ‘1.8Hz’ and ‘2.6Hz’ (Fig. 6a). Each chunk was synthesized as a 2-digit unit, using the AT&T Text-to-Speech System accentuation¹¹. To enable the generation of stimuli with chunk rates greater than the delta frequency upper bound (at 2.6 Hz), the waveforms in all conditions were time compressed by a factor of 2.5, just below the auditory channel capacity (Ghitza, 2014). Note that a particular 2-digit chunk has the same acoustics, regardless of whether it occurs in the 1.8Hz or 2.6Hz 2222 chunk condition (Fig. 6a, red box). The 1.8Hz stimulus is generated by increasing the gap between the chunks (with identical chunk acoustics). Overall, we used 200 10-digit stimuli (100 stimuli for each of the 1.8Hz and 2.6Hz conditions); for each stimulus, a trial was created by concatenating the following waveform sequence: [one digit trial-count] [20-msec long gap] [10-digit string] [500-msec long gap] [2-digit target], resulting in one concatenated waveform per trial with durations that varied across trials and conditions. The 200 trials were scrambled, and the resulting pool of trials was divided into blocks, 50 trials per block.

Task. Behavioral and MEG data were collected while participants performed a digit retrieval task, in the form of an adapted Sternberg target identification task (Sternberg, 1966) (target ID task from here on): listeners heard a 10-digit stimulus followed by a 2-digit long target, and were asked to indicate whether or not the target was part of the utterance. A target position was always within

¹¹ The AT&T-TTS system (<http://www.wizzardsoftware.com/text-to-voice.php>) uses a form of concatenative synthesis based on a unit-selection process, where the units are cut from a large, high-quality, pre-recorded natural voice fragments. The system produces natural-sounding, highly intelligible spoken material with a realistic prosodic rhythm—with accentuation defined by the system-internal prosodic rules—and is considered to have some of the finest quality synthesis of any commercial product.

a chunk. Note that the task is suitable for probing the role of acoustic segmentation in a memory retrieval task: a successful yes/no decision depends on how faithful the recognized chunk objects are, generated by a decoding process that, by hypothesis, depends on the goodness of segmentation.

Behavioral Analysis. A “yes–no” model for independent observations was used to compute *dprime* (Green and Swets, 1966). Four classes of response were considered: (1) Hit: a “yes” response when the target chunk is present in the digit sequence, (2) Correct Rejection: a “no” response when the target chunk is absent, (3) Miss: a “no” response when the target chunk is present, and (4) False Alarm: a “yes” response when the target chunk is absent. Wilcoxon signed-rank tests (two-sided) were used to test differences in the mean *dprime* across conditions.

Neuroimaging Analysis. The scientific/technical detail involved in the MRI and MEG data acquisition, the MRI analysis, the MEG preprocessing, and the MEG source-level analysis [which map the sensors onto voxels, defined across cortical Regions Of Interest (ROIs, see Fig. 6b)] can be found in the publication. The brain waves (one per voxel) generated by this sequence of operations are the input to the periodicity density analysis described next.

Periodicity density function (PDF) within ROI. We aimed to determine whether the elicited brain signal measured at any given voxel within a specific ROI shows periodicity, and if so, to extract the frequency. Ultimately, we seek to characterize the density function of the periodicities across all voxels in the ROIs of interest. The PDF within ROI is obtained in two steps. First, for each voxel, periodicity is computed using a newly proposed measure method termed XCOV. The desired PDF is the histogram of the XCOV periodicities across all voxels in the prescribed ROI. A detailed description of XCOV can be found in subsection 3.4 (Analysis tools) below.

Behavioral results. The *dprime* values are shown in Fig. 6c, separately for each condition. Blue dots indicate individual *dprime* scores, black lines indicate the mean *dprime* scores, dark gray bars indicate the ± 1 standard error of the mean and light gray bars the confidence interval. Significance is indicated by **($p < .01$). The performance was higher in the 1.8Hz acoustical chunk (inside delta chunking), compared to the 2.6Hz acoustical chunk condition (outside of delta), replicating findings in Ghitza (2017).

Neuroimaging results. Our findings show that prosodic chunking of phrase-sized acoustic segments is correlated with acoustic-driven delta oscillations, expressing anatomically specific patterns of neuronal periodicities. The data show that in superior and middle temporal areas and in speech-motor planning and integration areas (IFG, PC, SMG), robust neural delta periodicities were elicited by chunk rates inside of the delta range but were diminished when the chunk rate was outside of the delta range, in line with behavioral performance. In speech-motor integration areas (SMG) and areas implicated in processing word form and meaning (MTG), periodicity was present albeit diminished even for chunk rates inside the delta range. Figure 6d shows examples of periodicity PDFs in three ROI regions (STG, MTG, IFG) elicited by stimuli inside (top) and outside (bottom) the delta range. In sum, the delta periodicities were acoustically driven, in sync with the input as long as the acoustic chunk rate was inside delta. Delta periodicities were diminished for chunk rates outside delta, even though sufficient acoustic cues for chunking were present in all conditions. Thus, the failure to track the input-chunk-rate when it was outside of the delta range was not caused by insufficient acoustic cues but seems due to neuronal circuitry characteristics constraining the tracking of the chunks.

3.2.2 Electrocorticographic responses to time-compressed speech vary across the cortical auditory hierarchy.

Davidesco I, Thesen T, Honey CJ, Melloni L, Doyle W, Devinsky O, Ghitza O, Schroeder C, Poeppel D, Hasson U (2018) Electrocorticographic responses to time-compressed speech vary across the cortical auditory hierarchy. *bioRxiv*. <https://doi.org/10.1101/354464>

Human listeners understand spoken language across a variety of speech rates, but when speech is presented at a rate faster than about three times its original rate, it becomes unintelligible. How the

brain achieves such tolerance and why speech becomes unintelligible above certain rates is still unclear. In this study we addressed these questions using electrocorticography (ECoG) recordings in 7 epileptic patients (two female). Patients rated the intelligibility of sentences presented at the original rate (100%), speeded rates (33% or 66% of the original sentence duration) and a slowed rate (150%). We then examined which parameters of the neural response covary with the transition from intelligible to unintelligible speech. Specifically, we asked whether neural responses: 1) track the acoustic envelope of the incoming speech; 2) “scale” with speech rate, i.e., whether neural responses elicited by slowed and speeded sentences can be linearly scaled to match the responses to the original sentence.

Participants. Seven native speakers of English (2 female; 24-56 years old) experiencing pharmacologically refractory complex partial seizures were recruited via the Comprehensive Epilepsy Center of the New York University School of Medicine. Patients had elected to undergo intracranial monitoring for clinical purposes and provided informed consent in accordance with New York University Medical Center Institutional Review Board. Electrode placement was determined based on clinical criteria without reference to this study. Patients had left-hemisphere (n=3), right-hemisphere (n=3) and bilateral (n=1) electrode coverage.

Stimuli. A set of 33 spoken sentences with a minimum duration of 3s were selected from the Harvard sentences corpus (IEEE, 1969). All sentences were recorded by a male speaker. Stimuli covered four speech rates: uncompressed (100%) speech, slowed (150%) and speeded speech (33% and 66% of the duration of the corresponding uncompressed signal; Fig. 7a). Sentences were presented consecutively, in pseudorandom order, until each sentence had been presented twice. To control for sentence duration across speeds, for the 33% condition we generated concatenated version (33C): sentences were generated by concatenating three different sentences first, and then time compress the concatenated group by a factor of 3 (Fig. 7a). (Thus, each of these speeded sentence-groups had the same duration as one of the original sentences.) The 8 sentences used to generate the 33C condition were different than the ones used in the other conditions and were sampled independently from the Harvard sentence corpus. Ten 33C sentences were generated in this manner, and each one of them was presented twice throughout the experiment, interleaved with the other conditions. Compression and dilation were performed by using the Overlap-Add algorithm in Praat (Boersma & Weenink, 2016).

Procedure. Participants listened to a total of 252 sentences, divided into two blocks. Sentences were played at bedside by a laptop and speakers located in front of the patient. Sentences were presented in a pseudo-random order, under the constraint that the same sentence was never repeated consecutively. The experiment was self-paced: following each sentence, patients verbally rated the intelligibility of the sentence they had just heard, using a 5-point scale from 1 (“not intelligible at all”) to 5 (“fully intelligible”).

Neuroimaging data acquisition. The technical detail involved in the ECoG Electrode localization, the ECoG signal acquisition, and the ECoG signal preprocessing can be found in the publication.

Neuroimaging Analysis. To compute the time course of the broadband power Morlet wavelets (standard deviation 6 cycles) with center frequencies at 70, 75, 80, ..., 200 Hz were convolved with the voltage time series. Amplitude time series at line-noise frequencies of 120 and 180 Hz were discarded, leaving 25 distinct time series. Each individual amplitude time series was logarithmically transformed and then converted to a z-series by subtracting its mean and dividing by its standard deviation. The high frequency broadband power was then estimated as the mean of all 25 of the z-series and smoothed with a hamming window of 125ms (Honey et al., 2012).

Statistical analysis, Neural tracking. Conducted for each electrode and for each speech rate. Neural tracking was defined as the correlation between the speech envelope of a sentence and the corresponding broadband ECoG response. The speech envelope was extracted for each sentence as follows: First, each sentence was filtered into sixteen critical bands logarithmically spaced between 230 and 3800 Hz; second, the Hilbert envelope was extracted for each band and then summed

across bands. Finally, the resulting envelope time-course was down-sampled to 512 Hz to match it to the ECoG signal (Doelling et al. 2014). The ECoG broadband response was first shifted backwards by the response latency of each electrode (as estimated from the external localizer – see below). Then, to reduce any components that are not sentence-specific, the mean neural response across all sentences of a given speech rate was regressed out of each trial. To account for the difference in signal length across compression/dilation conditions, both the sentence envelope and the ECoG broadband responses were resampled to match the original sentence duration (i.e. 33% and 66% responses were up-sampled, 150% responses were down-sampled). Next, the first 300 ms and the last 300 ms were cropped from each trial in order to exclude onset or offset-related transients. Finally, the ECoG responses were averaged across the two repetitions of each sentence and correlated with the sentence envelope. Note that a correlation analysis was used, rather than a phase-locking analysis, because the latter requires many repetitions of each sentence, which would have influenced the intelligibility of compressed speech.

Statistical analysis, Linear scaling. This analysis was used to test the extent to which linear scaling of the neural responses elicited by speeded or slowed down sentences match the neural responses to the original speech rate. In this analysis, the response to the original sentence (100%) was always used as a reference signal, and the neural response to each one of the other speech rates was resampled to match the original sentence duration (responses to speeded speech were up-sampled, responses to slowed speech were down-sampled; Lerner et al., 2014). Then, for every speech rate, the resampled neural response was correlated with the response to the original sentence. Note that the linear scaling analysis is expected to provide additional information, not captured by the speech tracking analysis. The speech envelope mainly reflects low-level properties of the speech signal (i.e. variations in amplitude over time). High-order cortical regions may no longer track the audio envelope of speech, but still be directly involved in speech processing (e.g., analyzing the grammatical structure of a sentence) (Honey et al., 2012). Therefore, linear scaling might be a more suitable measure to compare neural responses across low- and high-level cortical areas.

Behavioral results (Fig. 7b). Patients rated the intelligibility of each sentence on a scale from 1 to 5 (where 5 is fully intelligible). Intelligibility was near ceiling for the 66%, 100% and 150% rates and, as expected, dropped sharply for the 33% and 33C conditions. Specifically, intelligibility significantly decreased from a level of 4.88 ± 0.05 (mean rating $\pm$ standard error of the mean) for the original duration to a level of 2.78 ± 0.38 for 33%, and to a level of 2.19 ± 0.34 for 33C.

Speech envelope tracking, results. We first assessed the extent to which neural responses tracked the audio envelope of each sentence. Figures 7c-A and 7c-B depict, for two different electrodes—low-level and high-level—the broadband neural responses (green) and audio envelope (red) for a single sentence at different speech rates. The low-level auditory electrode (Fig. 7c-A) closely tracked the audio envelope for all speech rates. In contrast, the high-level STG electrode (Fig. 7c-B) showed a sharp decrease in envelope tracking for the 33% and 33C conditions. To investigate the spatial topography of neural tracking of the speech envelope, we conducted a whole-brain analysis by assessing the significance of speech tracking across all recorded electrodes. Figure 7c-C shows the speech envelope tracking maps for each speech rate. Even though we did not have access to neural data from Heschl’s gyrus, seven electrodes – localized mainly along the lateral sulcus in the vicinity of early auditory areas – displayed significant speech tracking for the most compressed speech levels (33% and 33C). For intelligible speech rates (66% and above), as expected, most of the electrodes that showed significant speech tracking were clustered along the STG as well as in the inferior frontal gyrus. Slowed-down speech (150%) yielded the highest number of speech tracking sites, with some extending to the angular gyrus and supramarginal gyrus.

Linear scaling, results. Here, we measured the extent to which the responses for the speeded (dilated) sentences match the original (100%) by up sampling (down sampling) the neural responses. Figure 7d depict the scaled neural responses (blue) and the response to the original sentence that served as a reference (pink). For example, in the case of 150% (slowed) speech, the

neural response was compressed (i.e., down sampled) to match the 100% response, and then the two responses were correlated. Like the speech tracking analysis, we identified two representative electrodes. For low-level auditory electrodes (Fig. 7d-A), significant response scaling was observed across all speech rates, even outside the intelligibility range. In contrast, for high-order STG electrodes (Fig. 7d-B), temporal scaling was observed only for intelligible speech (66% and 150%) and not for unintelligible speech (33% and 33C). This step-like transition in temporal scaling from intelligible to non-intelligible speech was also evident in a whole-brain analysis. Significant temporal scaling along STG, Inferior Frontal Gyrus (IFG) and supramarginal gyrus was observed for intelligible speech (66% and 150%). In contrast, scaling of neural responses to unintelligible speech was mainly confined to STG sites near early auditory cortex. Finally, Fig. 7d-C shows the spatial distribution of the electrodes with a significant temporal scaling, for each speech rate.

In sum, our data show that intelligibility was at ceiling for speech rates of 66% and slower but dropped significantly for the 33% rate. At the neural level, STG regions near A1 ('low-level') tracked the acoustic envelope and linearly scaled with the input across all speech rates, irrespective of intelligibility. In contrast, secondary auditory areas in the STG as well as the inferior frontal gyrus and angular gyrus ('high-level') tracked the acoustic envelope and linearly scaled with input only for intelligible speech. These results provide better understanding of how information processing unfolds along the cortical auditory hierarchy.

3.2.3 Dynamics of functional networks for syllable and word-level processing.

Rimmele JM, Sun Y, Michalareas G, Ghitza O, Poeppel D (2019) Dynamics of functional networks for syllable and word-level processing. *Neurobiology of Language* (under review). *bioRxiv*.
<https://doi.org/10.1101/584375>

The comprehension of speech requires integrating information at multiple time scales, including phonemic, syllabic, and word scales. Typically, we think of decoding speech in the service of recognizing words as a process that maps from phonemic units to words. Recent neurophysiological evidence, however, has highlighted the relevance of syllable-sized chunks for segmenting speech through neuronal oscillatory tracking in auditory cortex (e.g., Ghitza, 2011 & 2013; Doelling et al., 2014). In this study we examine how syllabic processing interacts with higher levels of speech processing, beyond segmentation, including the anatomical and neurophysiological characteristics of the networks involved. In two MEG experiments, we investigate the effects of lexical processing and the interactions with (acoustic) syllable processing using a frequency-tagging paradigm. Participants listened to disyllabic words presented at a rate of 4 syllables/sec (hence, a rate of 2 words/sec). Lexical content (native language), sub lexical syllable transitions (foreign language), or mere syllabic information (pseudo words) were presented. Two conjectures were evaluated: (i) lexical processing of words activates brain areas that interact with (single) syllable processing; and (ii) beyond lexical processing, syllable transitions contribute to word-level processing.

Schematic description of the paradigm and material. A schematic description of the paradigm is shown in Fig. 8a. Participants were asked to listen to sequences of disyllabic words, that contained either German, Turkish, or Turkish pseudo-words. Stimuli were generated by a German/Turkish female bilingual speaker. Listeners were German natives with no previous knowledge of Turkish. In Experiment 1, a German condition and a Turkish condition were used. In the German condition (Fig. 8a-B), syllables (presented at a rate of 4 Hz, 250-ms purple bar) can be grouped (by the German native listener) into disyllabic words (black: first syllable, gray: second syllable) at a rate of 2 Hz (500-ms blue bar). In the Turkish condition (Fig. 8a-C), syllables (presented at 4 Hz) cannot not be grouped (by a *German* native listener) into words based on semantic knowledge. In Experiment 2, sequence of German words (Fig. 8a-B) and Turkish pseudo-words (Fig. 8a-D) were used. Each sequence contained 38 syllables that formed 19 'disyllabic units' (word or pseudo-words). By presenting isochronous syllables, the presentation rate for syllables was fixed

at 4 Hz, resulting in an occurrence of disyllabic units at 2 Hz. In order to maintain participants' attention on the auditory stimuli, a target stimulus, which consisted of a disyllabic unit that was made of a repetition of the same syllable, was inserted in 28.57% of the trials. Each block contained 29% trials with a target stimulus (equally distributed across conditions). Participants were asked to indicate by a button press, at the end of each trial (question mark icon in Fig. 8a-A) whether a target stimulus (in bold letters) was present. Each trial was followed by a jittered intertrial interval (2 - 3.5 sec).

Neuroimaging Analysis. The scientific/technical detail involved in the MRI and MEG data acquisition, the MRI analysis, the MEG preprocessing, and the MEG source-level analysis [which map the sensors onto voxels, defined across cortical Regions Of Interest (ROIs, see Fig. 6b)] can be found in the publication. The brain waves (one per voxel) generated by this sequence of operations are the input to the periodicity density analysis described next.

Neuroimaging Results. The left column of Figure 8b-A to 8b-D shows power contrasts (with the neighboring frequency bins) averaged across the individual M100 sensors. The right column shows the topography of the power contrast differences between conditions at 2 Hz (word level) and 4 Hz (syllable level); clusters that show significant differences are marked with black dots. In Experiment 1, Lexical processing resulted in increased power at 2 Hz for the German compared to the Turkish condition (Fig. 8b-A). Syllable transition processing resulted in increased power in the Turkish compared to the Non-Turkish (pseudo-words) condition at 2 Hz (Fig. 8b-B). In Experiment 2, Lexical plus syllable transition processing resulted in increased power at 2 Hz in the German compared to the Turkish condition at a broadly distributed cluster (Fig. 8b-C). No differences were detected for the across experiment comparison of the German conditions (Fig. 8b-D). In summary, the data provide a new perspective on speech comprehension by demonstrating a contribution of an acoustic syllabic to lexical processing route. The data show that syllable-transition information, compared to mere syllable information, activated a bilateral superior-, middle temporal and inferior frontal network. Lexical content additionally activated left-lateralized frontal brain areas. Syllable tracking in auditory cortex (quantified by a cerebro-acoustic coherence measure, see publication) decreased when lexical content was present, indicating an interaction of word- and syllable-level processing. More specifically, right auditory cortex syllable processing interacted with word-level processing in right frontal, superior and middle temporal areas (cross-frequency coupling).

Finally, figure 8c depicts the interaction of syllable and word-level processing. In Fig. 8c-A, we propose to extend the classic 'Phoneme Route' models via contributions of a 'Syllabic Route' to word-level processing. In Fig. 8c-B, a schematic description of the interaction of syllable and word-level processing is shown. Frequency tracking of the syllabic rate in STG (4 Hz activity, top row) affects higher-level syllable-transition processing in a network including bilateral STG, MTG and IFG (2 Hz activity, middle row), and higher-level lexical processing in a network including STG and left MTG and IFG and interactions between STG and left middle temporal areas (2 Hz activity, bottom row). The interactions between syllable and word-level processing occur in STG and MTG (red line). The intensity of the gray shades indicates the activation level during the respective process.

3.3 Infrastructure

3.3.1 Contextually controlled stories Corpus

Lewis G, van Rijn P, Gwilliams L, Larrouy-Maestri P, Poeppel D, Ghitza O (2021) NyU-BU contextually constrained sentences Corpus: NUBUC. *Zenodo Data Repository*.

<https://zenodo.org/record/4075183#.YKVMoi0RmF>

To study the role of sentential context for speech decoding and ultimately test the neural correlates, developed a corpus of linguistically contextually constrained sentences, which provide additional control over the properties of spoken material, beyond the SUS (Semantically Unpredictable Sentences; Ghitza & Greenberg, 2009) and TIMIT. TIMIT is the gold standard of

spoken sentences. A key asset is the phonetically transcribed material, which simplifies and speeds error analysis. TIMIT's structure was developed to reduce the effect of semantic context – the sentences are somewhat anomalous (e.g., “She had your dark suit in greasy wash water all year”), but can be partially decoded with contextual information. The SUS corpus (e.g., “the school poured the camp that came”) was developed to reduce semantic cues even further. However, it still conforms to limited grammatical structures and is not entirely unpredictable. A third database that has been used extensively by the PI are digit strings (e.g. Ghitza 2012; Ghitza 2014); these have no syntactic context but, because of the limited target set of items, a low perplexity.

The success of a language experiment heavily relies on selecting stimulus materials with controlled linguistic properties of the spoken material. This selection process entails a critical tradeoff between similarity to ‘real’ language (i.e., external validity) and experimental and analytic control (i.e., internal validity). In order to bridge these conflicting demands, we developed the NyU-BU contextually controlled stories Corpus (NUBUC) of spoken language. The corpus is both naturalistic and experimentally controlled, comprising 16 high-quality recordings of 8 unique stories, spoken both by a female and a male actor. Each story consists of 128 sentences (~2000 words per story) organized around critical keywords, which have been matched along multiple linguistic dimensions. The context surrounding each keyword is also parametrically manipulated, varying prior context (weak/strong), local context (weak/strong) and sentence position (early/late). Here we describe the corpus in detail, including how it compares to and builds on existent corpora. These materials showcase the ability to overcome the apparent dichotomy between control and generalizability, by presenting subjects with carefully curated linguistic materials in a naturalistic listening scenario. The corpus is freely available. Overview of corpus is shown in Fig. 9a. (A) Each pair of sentences contains a target word (or keyword), a local context word, and a prior context word or a phrase. The keyword and the local context always appear in the second sentence and the prior context in the first sentence. The corpus controls four variables: syntactic position, prior context, local context, and the target word. (B) Target words are part of a semantic category, have a given number of syllables and are grouped into low and high frequency targets. (C) Local context describes the probability of the keyword following the local context word. The bigram of the local context and keyword is divided by the unigram of the local context word (Ngrams from CoCa). All sentence pairs with an above average ratio have a strong context, all pairs below average have a weak context. (D) Prior context describes the probability of the prior context word occurring shortly before the target word. It is strong if it is plausible, it is weak if it is not plausible. (E) The syntactic position refers to the local context and keyword located in the second sentence, so early or late in the sentence. (F) Each story contains 64 sentence pairs. With 8 stories this yields 1.024 sentences. Each story contains all possible condition combinations 8 times. Fig. 9b shows a demo display of the waveforms and part of accompanied meta data of sentence pair 1, story 1, male speaker. The prior context (yellow), local context (cyan) and the target word (magenta) are highlighted. Below the waveform there is the automatically parsed syntactic tree. The syntactic trace is indicated with a line.

3.4 Analytic tools

3.4.1 Periodicity density function (PDF) within cortical Region Of Interest (ROI).

Rimmele JM, Poeppel D, Ghitza O (2021). "Acoustically driven cortical delta oscillations underpin prosodic chunking." *eNeuro*. <https://doi.org/10.1523/ENEURO.0562-20.2021>

Neuroimaging studies on the role of neuronal oscillators in cortical computation circuitry require the determination whether the elicited brain signals measured at any given voxel within a specific ROI show periodicity, and if so, to extract the frequency. Ultimately, we seek to characterize the density function of the periodicities across all voxels in the ROIs. The analysis pipeline for deriving the periodicity density function comprises two steps: (i) measuring the neural response periodicity

in individual voxels, and (ii) The derivation of the periodicity density function (PDF) within the ROI.

(i) The aggregated cross-correlation measure (XCOV) of periodicity. To measure the neural response periodicity in individual voxels one could use one of several widely used measures, e.g., autocorrelation—where the first nontrivial peak indicates the period, or the intertrial phase coherence (ITPC)—where the outcome would be the frequency distribution of the coherence function. Importantly, these measures build on the number of trials, M . The trial signals are noisy, both due to the SNR and due to the brain wave irregularity (which is why these methods average over trials). But what if M is too small? Here we used a newly proposed measure, termed ‘Aggregated cross-correlation’ (abbreviated XCOV), to measure periodicity across M trials. Broadly, we suggest taking advantage of the fact that, for M trials, we can generate about $M^2/2$ cross-correlation functions. Recall that, unlike autocorrelation, the first peak of a cross-correlation function does not indicate the period but rather the delay between the two signals. Therefore, we run each of the $M^2/2$ candidate pairs through a “match filter”, which determines whether the corresponding two signals have a “zero” delay. Such a pair will have a cross-correlation function similar to that of an autocorrelation function, i.e., its peak is at zero and its earliest nontrivial peak is at the period. Only the pairs that pass the test are cross-correlated and aggregated. Obviously, the number of cross-correlation functions qualified for aggregation is between M and $M^2/2$, depending on how strict the match filter is. (For example, in the STG ROI, the mean number of trials over subjects for the ‘Hit’ response was $M=38$, with a mean number of pairs of 703. The mean number of pairs that passed the test was 433 for the 1.8Hz condition and 378 for the 2.6Hz condition – about one degree of order bigger than M . A similar trend was observed for all ROIs.) We term the outcome of this measure as the ‘XCOV’ function.

(ii) Periodicity density function (PDF) within a ROI. Figure 10 details the analysis pipeline for deriving the periodicity density function of the periodicities within a particular ROI. L voxels, N subjects, and M trials per subject are considered. First (not shown), each brain signal is filtered to the frequency range of interest. Cross correlations were computed using the filtered signals. Shown is the XCOV function at the i -th voxel, for the j -th subject, obtained by aggregating K cross-correlation functions. (Note that as a cross-correlation function, XCOV is computed against time-lags; the abscissa here shows the time-lag inverse, in frequency, hence going from right to left). The particular XCOV function shown in Fig. 10 has a single peak at 1.76 Hz but note that, in general, an XCOV may have more than one local peak. Next, the location of the prominent peaks is extracted, with the number of prominent peaks as a parameter. (The prominence of a peak measures how much the peak stands out due to its intrinsic height and its location relative to other peaks in the range of interest.) In our analysis one prominent peak per XCOV is considered. Hence, for L voxels and N subjects, a maximum of $L \times N$ data points are available to construct a histogram, from which only those inside the frequency range of interest are used, and the resulting histogram is normalized to $L \times N$. A 3rd order Gaussian mixture model (GMM) that fits the histogram is the desired PDF. The “goodness” of the periodicity is quantified by in terms of P , the percentage of datapoints inside the frequency range of interest with respect to the total number of datapoints ($L \times N$); and the mean μ and variance σ of the prominent Gaussian component of a 3rd order GMM. (The total number of data points is shown in the inset of each entry.)

3.4.2 The modulation-spectrum cube representation of speech. (Unpublished.)

The cube is a 3-dimention representation of speech defined by time, modulation frequency and cochlear place. It was developed with the purpose of revealing the role of prescribed modulations (oscillatory in nature) in decoding syllables and phrases. Figure 11 outlines the modulation-spectrum cube structure. The upper and middle panels of Fig. 11-A show the waveform and the Fourier wide-band spectrogram of a typical 2-sec long speech sentence, while the bottom panel shows the temporal envelopes of ten randomly selected cochlear channels, lowpass-filtered to 50

Hz. By definition, the modulation spectrum of a cochlear channel is the Fourier transform of its temporal envelope. As such, it comprises the sinusoidal components (modulations) of the temporal envelope—magnitude and phase. Fig. 11-B shows a 2-D modulation spectrum of a cochlear channel's temporal envelope at 2200 Hz, smoothed to 50 Hz. Shown are the half-wave rectified modulations in the [0.5–7.6 Hz] modulation frequency range. The 3-D cube, shown in Fig. 11-C, is the sequence of the 2-D modulation spectra (one per cochlear channel) evolving along the cochlear partition, hence a representation along the time, modulation frequency and cochlear place axes. (Alternatively, the 3-D cube can be displayed as a movie, where the 2-D modulation spectra—one per cochlear place along the cochlear partition—are the movie frames.) The modulation-spectrum cube characterizing the relationship between the syllabic and prosodic levels of speech perception and prescribed 3D sub-cubes. And since modulation spectrum at a given cochlear place is oscillatory in nature, these relationships will reveal the role of oscillation-frequency regions in decoding syllables and phrases.

3.5 Impact of research

It is no exaggeration to assert that the perspective we (the PI and David Poeppel) have jointly developed and advocated, as well as the results we have published (both singly and jointly) on these topics, have had significant influence on the literature on cortical speech processing and its neural foundations. For example, the tandem of two papers, the Ghitza (2011) paper on “Linking speech perception and neurophysiology” and the Giraud & Poeppel (2012) paper on “Cortical oscillations and speech processing” have been cited (Google Scholar) 282 times and 775 times, respectively. The foundational Ghitza & Greenberg (2009) psychophysics paper on time-compressed speech has been cited 274 times; and two papers in which we jointly test the causal role of theta-entrainment in a “manipulated modulation spectrum” (Ghitza, 2012; Doelling et al. 2014) have been cited 113 and 334 times, respectively; and a recent paper in which the role of delta-entrainment (Ghitza, 2017) has already been cited 46 times. Another paper, which extends our approach to the delta band and incorporates phrase- and sentence-level processing (Ding et al. 2016), has been cited 302 times. Cumulatively, it is fair to say that current work on spoken language processing across labs—domestic and abroad—is engaging in a serious manner with the ideas set forth in our work, and there are now hundreds of citations to the work we have published.

It is important to be clear that the research program we are pursuing has not just been received in a parochial manner in the speech literature but has penetrated other aspects of research. The conceptual model we have developed has been tested and expanded in recent experimental work on music perception (e.g., Farbood et al. 2014), engineering (e.g., Räsänen et al. 2015), studies of dyslexia (e.g., Goswami 2011; Lehongre et al. 2011), and even some work on dynamic aspects of visual perception (VanRullen et al. 2015).

References

- Ahissar E, Ahissar M (2005) Processing of the temporal envelope of speech. In *The auditory cortex. A synthesis of human and animal research*. (eds. Konig R, Heil P, Bundinger E, Scheich H) Ch 18. New-Jersey: Lawrence Erlbaum
- Ahissar E, Haidarliu S, Zacksenhouse M (1997) Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators. *Proc Natl Acad Sci USA* 94:11633–11638
- Assaneo MF, Rimmele JM, Sanz Perl Y, Poeppel D (n.d.) Speaking rhythmically can shape hearing. *Accept Nat Hum Behav*.
- Aubanel V, Davis C, Kim J (2016) Exploring the Role of Brain Oscillations in Speech Perception in Noise: Intelligibility of Isochronously Retimed Speech. *Front Hum Neurosci* 10:430.
- Auer P, Couper-Kuhlen Elizabeth, Müller Frank (1999) Language in time: the rhythm and tempo of spoken interaction. New York: Oxford University Press.
- Boersma P, & Weenink D (2016) Praat: doing phonetics by computer [computer program].
- Bonhage CE, Meyer L, Gruber T, Friederici AD, Mueller JL (2017) Oscillatory EEG dynamics underlying automatic chunking during sentence processing. *NeuroImage* 152:647–657.
- Boucher VJ, Gilbert AC, Jemel B (2019) The Role of Low-frequency Neural Oscillations in Speech Processing: Revisiting Delta Entrainment. *J Cogn Neurosci* 1–11.
- Breen M (2018) Effects of metric hierarchy and rhyme predictability on word duration in The Cat in the Hat. *Cognition* 174:71–81.
- Bröhl F, Kayser C (2020) Delta/theta band EEG differentially tracks low and high frequency speech envelopes. *bioRxiv* 2020.07.26.221838.
- Brungart D, Wassenhove V van, Brandewie E, Romigh GD (2007) The effects of temporal acceleration and deceleration on AV speech perception In: AVSP.
- Buiatti M, Peña M, Dehaene-Lambertz G (2009) Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage* 44:509–519.
- Chi T, Gao Y, Guyton M, Ru, P, Shamma, SA (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* 106:2719–2732
- Clifton CJ, Carlson K, Frazier L (2006) Tracking the what and why of speakers' choices: prosodic boundaries and the length of constituents. *Psychon Bull Rev* 13:854–861.
- Deniz ND, Fodor JD (2019) Timing of Syntactic and Rhythmic Effects on Ambiguity Resolution in Turkish: A Phoneme Restoration Study. *Lang Speech* 0023830919894614.
- Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19:158–164.
- Ding N, Simon JZ (2009) Neural Representations of Complex Temporal Modulations in the Human Auditory Cortex. *J Neurophysiol* 102:2731–2743.
- Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, 85:761–768
- Escudero P, Benders T, & Lipski SC (2009) Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, 37(4), 452–465.
- Farbood MF, Rowland J, Marcus G, Ghitza O, Poeppel D (2014) Decoding time for the identification of musical key. *Atten Percept Psychophys*. doi:10.3758/s13414-014-0806-0
- Fletcher H, Steinberg JC (1929) Articulation testing methods. *Bell System Technical Journal* 8:809–854.
- Foulke E, Sticht TG (1969) Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, 72:50–62
- French NR, Steinberg JC (1947) Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* 19:90–119.
- Garvey WD (1953) The intelligibility of speeded speech. *Journal of Experimental Psychology*, 45:102–108.
- Gee JP, Grosjean F (1983) Performance structures: A psycholinguistic and linguistic appraisal. *Cognit Psychol* 15:411–458.
- Ghitza O (2020) “Acoustic-driven oscillators as cortical pacemaker”: a commentary on Meyer, Sun & Martin (2019). *Lang Cogn Neurosci* 35:1100–1105.
- Ghitza O (2017) Acoustic-driven delta rhythms as prosodic markers. *Lang Cogn Neurosci* 32:545–561.
- Ghitza O (2014) Behavioral evidence for the role of cortical theta oscillations in determining auditory

- channel capacity for speech. *Front. Psychology*, 5:652. doi:10.3389/fpsyg.2014.00652
- Ghitza O (2013) The theta-syllable: a unit of speech information defined by cortical function. *Front. Psychology*, 4:138. doi:10.3389/fpsyg.2013.00138
- Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front. Psychology* 3:238. doi:10.3389/fpsyg.2012.00238
- Ghitza O (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front. Psychology* 2:130. doi:10.3389/fpsyg.2011.00130
- Ghitza O, Giraud AL, Poeppel D (2013) Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Front. Hum. Neurosci.* 6:340. doi:10.3389/fnhum.2012.00340
- 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* 66:113–126. doi:10.1159/000208934
- Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neurosci.* 15(4):511–517. doi:10.1038/nn.3063
- Gordon-Salant S, and Fitzgibbons PJ (1993) Temporal factors and speech recognition performance in young and elderly listeners. *J. Speech Hearing Res.* 36, pp. 1276–1285.
- Goswami U (2011) A temporal sampling framework for developmental dyslexia. *Trends Cogn Sci.* 15(1):3–10.
- Green DM, Swets JA (1966) Signal detection theory and psychophysics. New York, Ny: Wiley.
- Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S (2013) Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biol* 11(12): e1001752. doi:10.1371/journal.pbio.1001752
- Haegens S, Zion Golumbic E (2017) Rhythmic facilitation of sensory processing: a critical review. *Neurosci Biobehav Rev.*
- Hilton C, Goldwater M (2020) Linguistic syncopation: Meter-syntax alignment and its effect on sentence comprehension and sensorimotor synchronization.
- Hochstein S, Ahissar, M (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36 (5): 791-804.
- Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, Devinsky O, Doyle WK, Rubin N, Heeger DJ, Hasson U (2012) Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* 76:423-434.
- Hovsepyan S, Olasagasti I, Giraud A-L (2020) Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nat Commun* 11.
- IEEE (1969) IEEE Recommended Practice for Speech Quality Measurements. In: *IEEE*, pp 1-24.
- Inbar M, Grossman E, Landau AN (2020) Sequences of Intonation Units form a ~ 1 Hz rhythm. *Sci Rep* 10:15846–15846.
- Keitel A, Gross J, Kayser C (2018) Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biol* 16:e2004473.
- Kidd, GR (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4), 736-748.
- Kösem A, Bosker HR, Takashima A, Meyer A, Jensen O, Hagoort P (2018) Neural Entrainment Determines the Words We Hear. *Curr Biol* 28:2867-2875.e3.
- Lakatos P, Gross J, Thut G (2019) A New Unifying Account of the Roles of Neuronal Entrainment. *Curr Biol CB* 29:R890–R905.
- Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005) An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *J Neurophysiol* 94:1904–1911.
- Lerner Y, Honey CJ, Katkov M, Hasson U (2014) Temporal scaling of neural responses to compressed and dilated natural speech. *Journal of neurophysiology* 111:2433-2444.
- Levit M (2004) Spoken Language Understanding without Transcriptions in a Call Center Scenario. *Universitt Erlangen–Nürnberg*. PhD dissertation.
- Levit M, Noeth E, Gorin A (2002) Using EM-trained string-edit distances for approximate matching of acoustic morphemes. *Proc. of ICSLP, ISCA*.
- Luce, P. A. and McLennan, C. (2005). Spoken word recognition: the challenge of variation. In *The handbook of speech perception* (eds. Pisoni DB, Remez RE) 591-609. Blackwell Publishing
- Marslen-Wilson, WD (1987) Functional parallelism in spoken word-recognition. *Cognition*, 25: 71–102

- Martin AE, Doumas LAA (2017) A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biol* 15:e2000663.
- Martin P (2015) *The Structure of Spoken Language: Intonation in Romance*. Cambridge: Cambridge University Press.
- Messing DP, Delhorne L, Bruckert E, Braidia LD, Ghitza O (2009) A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise. *Speech Communication* 51:668–683. doi:10.1016/j.specom.2009.02.002
- Meyer L, Gumbert M (2018) Synchronization of Delta-Band Oscillations to Syntax Benefits Linguistic Information Processing. *Journal of Cognitive Neuroscience*, 30:1066-1074. doi:10.1162/jocn_a_01236
- Meyer L, Henry M, Gaston P, Schmuck N, Friederici A (2016) Linguistic Bias Modulates Interpretation of Speech via Neural Delta-Band Oscillations. *Cereb Cortex* 27.
- Miller GA (1962) Decision Units in the Perception of Speech. *IRE Trans. Information Theory*, Feb., pp. 81–83.
- Miller GA (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Moulines E, Charpentier F (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.
- Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci Adv* 5:eaay6279
- Peelle JE, Davis MH (2012) Neural oscillations carry speech rhythm through to comprehension. *Front. Lang. Sci.* 3:320. doi:10.3389/fpsyg.2012.00320
- Pickett, JM, & Decker, LR (1960) Time factors in perception of a double consonant. *Language and Speech*, 3(1), 11-17.
- Pittman-Polletta BR, Wang Y, Stanley DA, Schroeder CE, Whittington MA, Kopell NJ (2020) Differential contributions of synaptic and intrinsic inhibitory currents to parsing via excitable phase-locking in neural oscillators. *bioRxiv*. doi:/10.1101/2020.01.11.902858
- Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time. *Speech Communication* 41:245–255. doi:10.1016/S0167-6393(02)00107-3
- Räsänen O, Doyle G, Frank MC (2015) Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *INTERSPEECH*. pp. 3204–3207
- Reinisch E, & Sjerps MJ (2013) The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, 41(2), 101-116.
- Rimmele JM, Poeppel D, Ghitza O (2021). "Acoustically driven cortical delta oscillations underpin prosodic chunking." *eNeuro*.
- Rimmele JM, Morillon B, Poeppel D, Arnal LH (2018b) Proactive Sensing of Periodic and Aperiodic Auditory Patterns. *Trends Cogn Sci* 22:870–882.
- Ryan J (1969) Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, 21:2, 137--147.
- Shamir M, Ghitza O, Epstein S, Kopell N (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. *PLoS Comput. Biol.* 5, e1000370. doi:10.1371/journal.pcbi.1000370
- Shattuck-Hufnagel S, Turk AE (1996) A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res* 25:193–247.
- Stehwien S, Meyer L (2021) Rhythm Comes, Rhythm Goes: Short-Term Periodicity of Prosodic Phrasing.
- Sternberg S (1966) High-speed scanning in human memory. *Science* 153, 652--654. doi: 10.1126/science.153.3736.652
- Stevens K (2005) Features in speech perception and lexical access. In *The handbook of speech perception* (eds. Pisoni DB, Remez RE) 125-155. Blackwell Publishing
- Toscano JC, & McMurray B (2015). The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience*, 30(5), 529-543.
- Vagharchakian L, Dehaene-Lambertz G, Pallier C, and Dehaene S (2012) A temporal bottleneck in the language comprehension network. *J. Neurosci.*, 32(26), pp. 9089-9102.
- VanRullen R, Zoefel B, Ilhan B (2014). On the cyclic nature of perception in vision versus audition. *Phil. Trans. R. Soc. B*, 369 (20130214).
- Viterbi AJ (1966) *Principles of coherent communication*. New York: McGraw-Hill.

- Wingfield A, Tun PA, Koh CK, and Rosen MJ (1999) Regaining lost time: adult aging and the effect of time restoration on recall of time-compressed speech. *Psychol. Aging* 14, pp. 380–389.
- Zacksenhouse M, Ahissar E (2006) Temporal decoding by phase-locked loops: unique features of circuit-level implementations and their significance for vibrissal information processing. *Neural Computation* 18:1611–1636.

Figure 1

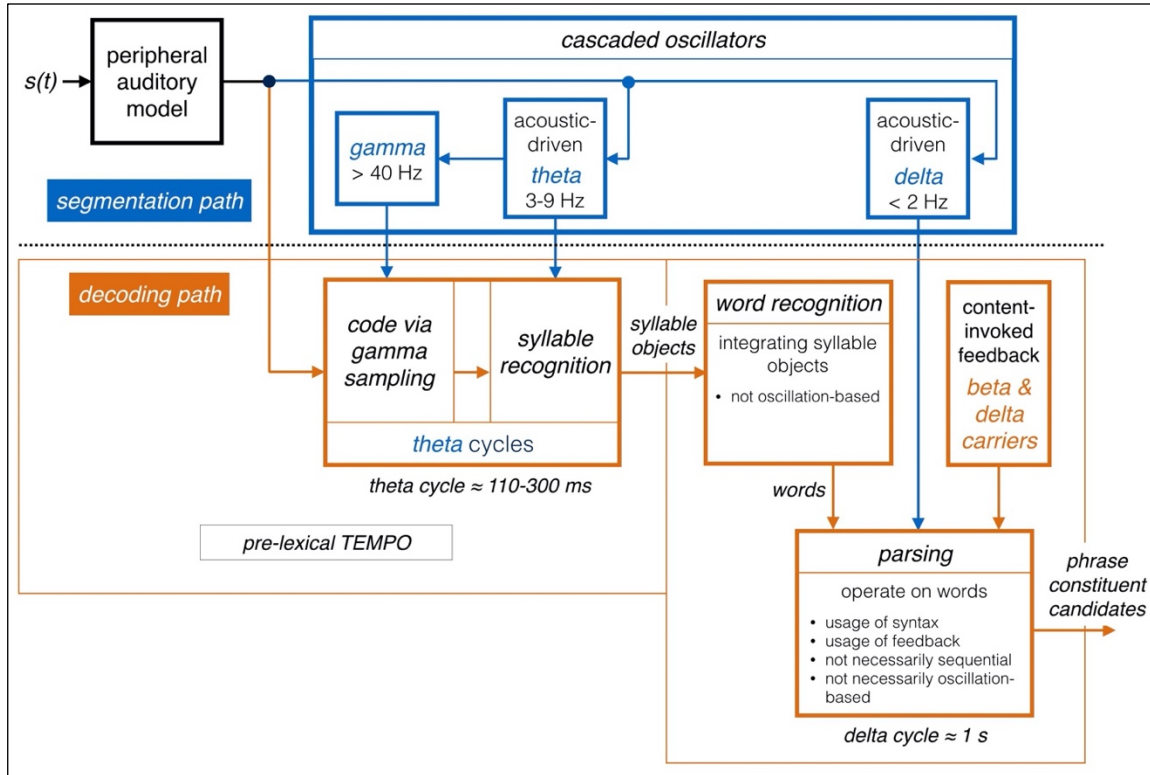


Figure 2a

Sequence of operations – syllable level

- 1 segmentation
 - window = a theta cycle
 - theta = a VCO in a PLL circuitry
 - evolving phase-locking process
- 2 neural code (evolves with theta)
 - gamma nested in theta
- 3 transmitting the code
 - end of theta cycle
- 4 by the end of next theta cycle
 - recognizing the code (to syllable object)
 - generating neural code for next chunk

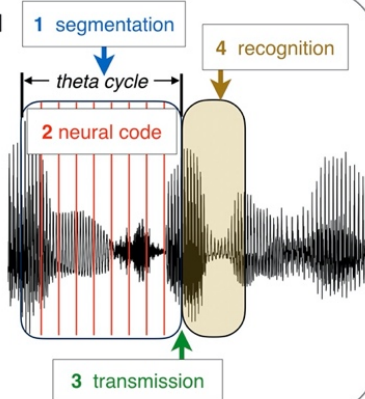


Figure 2b

Sequence of operations – phrase level

- 1 segmentation
 - window = a delta cycle
 - delta = a VCO in a PLL circuitry
 - evolving phase-locking process
- 2 parsing (evolves with delta)
 - generating PCCs
 - operates on words
 - not necessarily sequential
 - content-invoked feedback
- 3 multiplexing the phrase candidates

Figure 2c

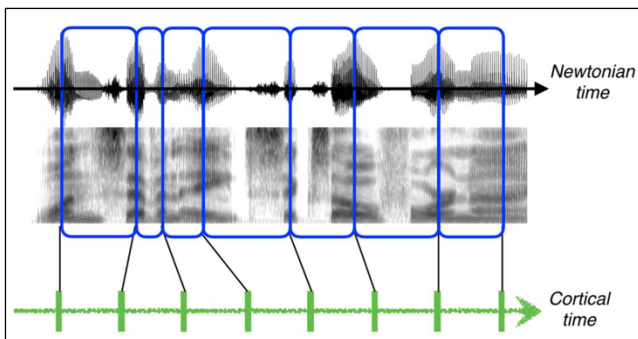


Figure 3a

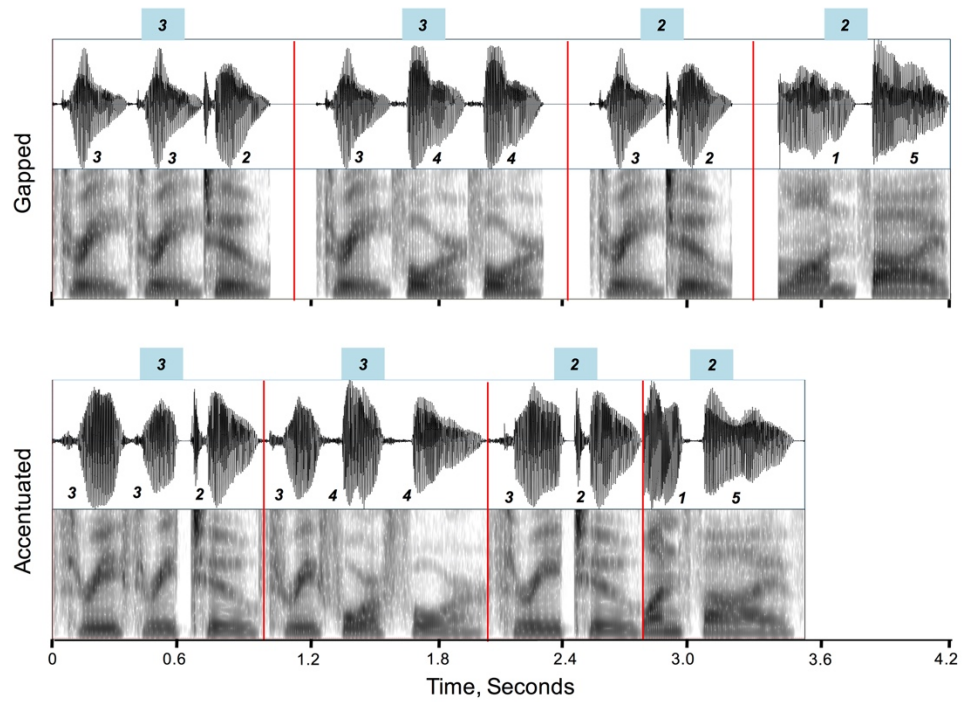


Figure 3b

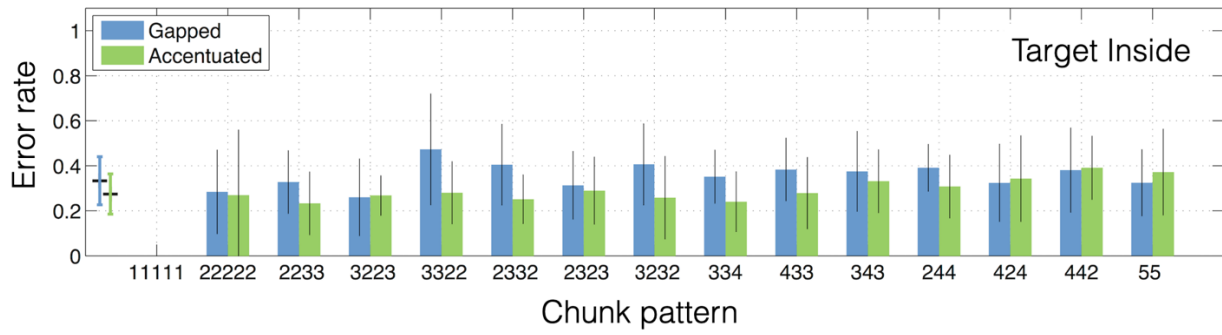


Figure 3c

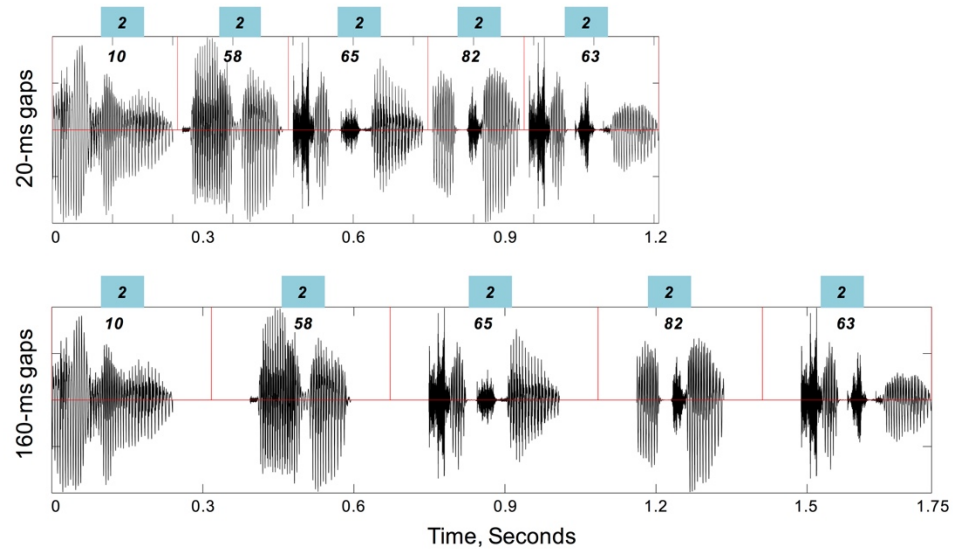


Figure 3d

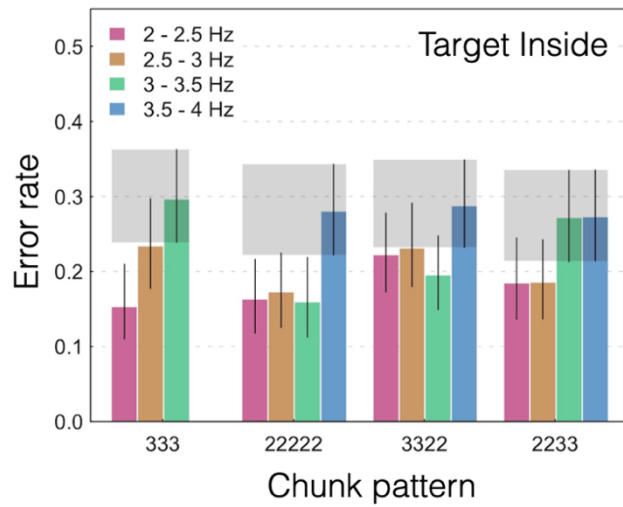


Figure 4a

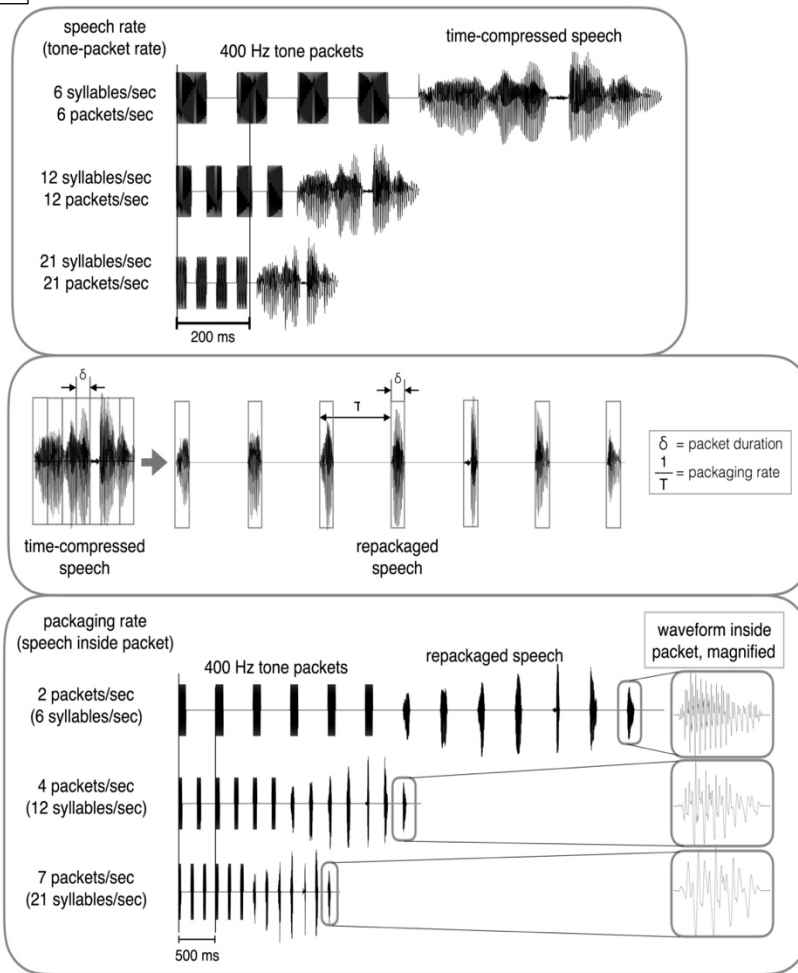


Figure 4b

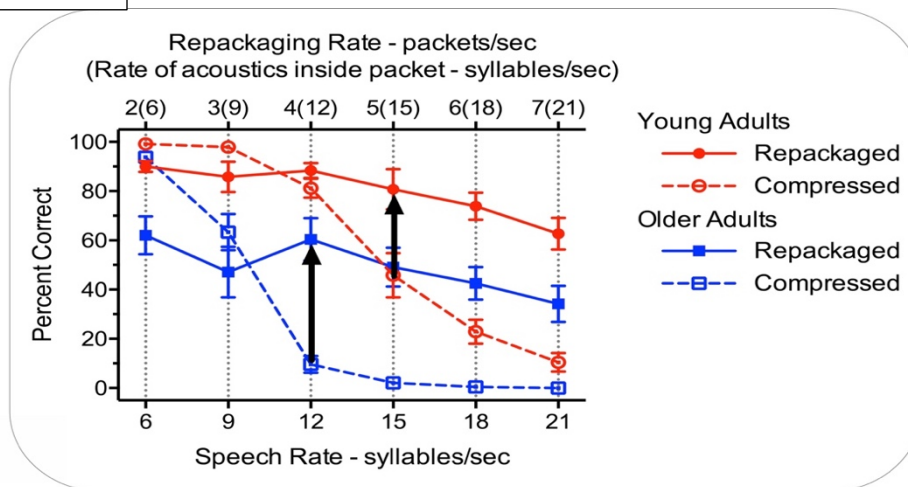


Figure 5a

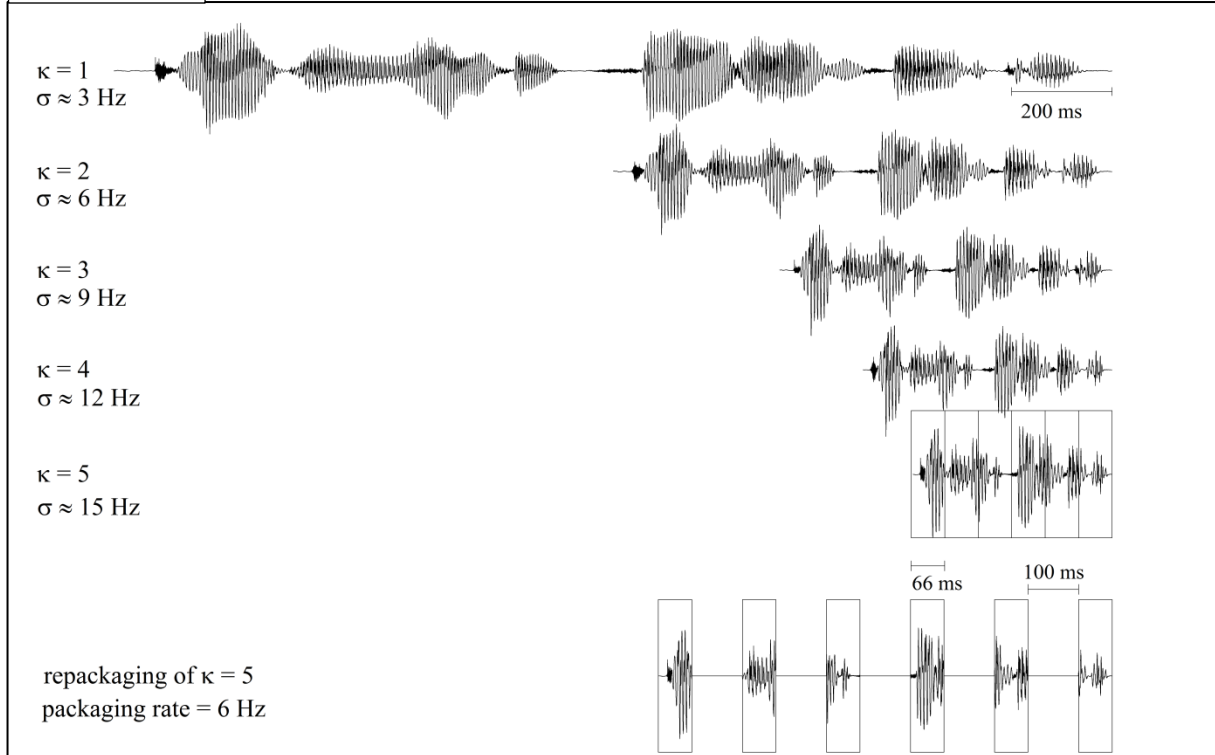


Figure 5b

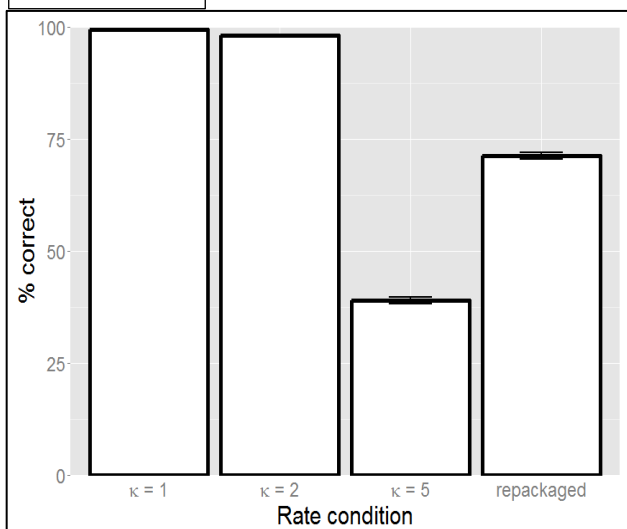


Figure 5c

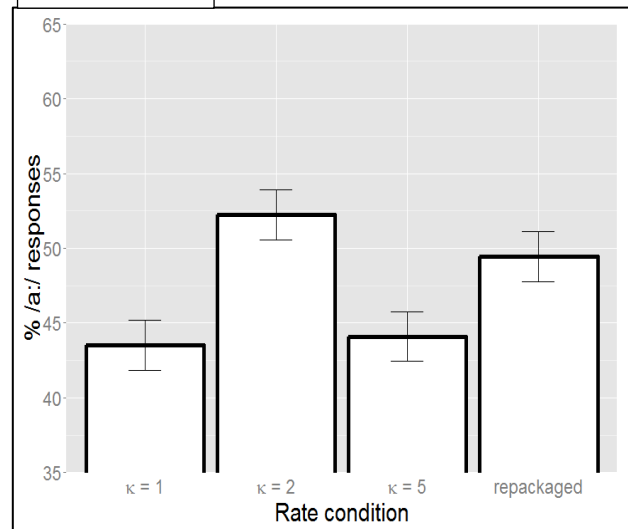


Figure 6a

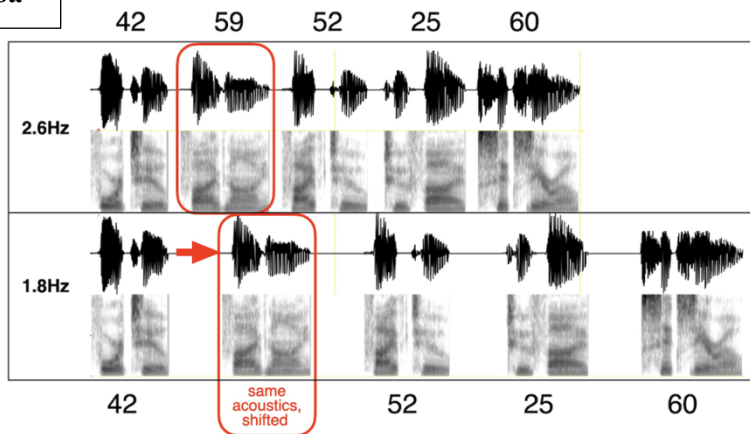


Figure 6b

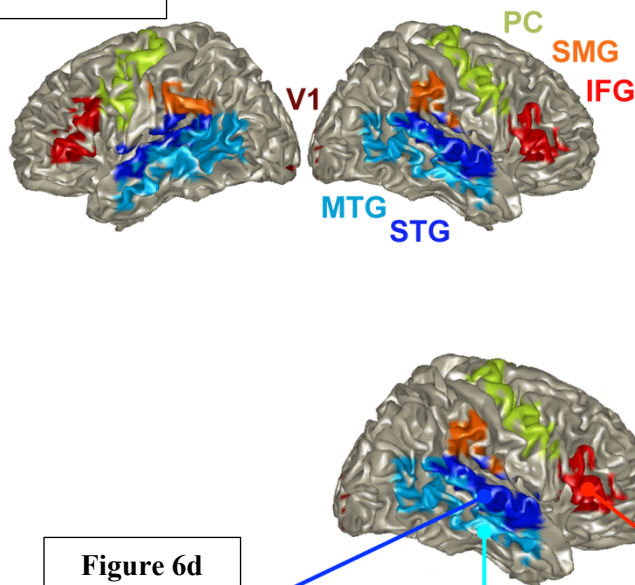


Figure 6d

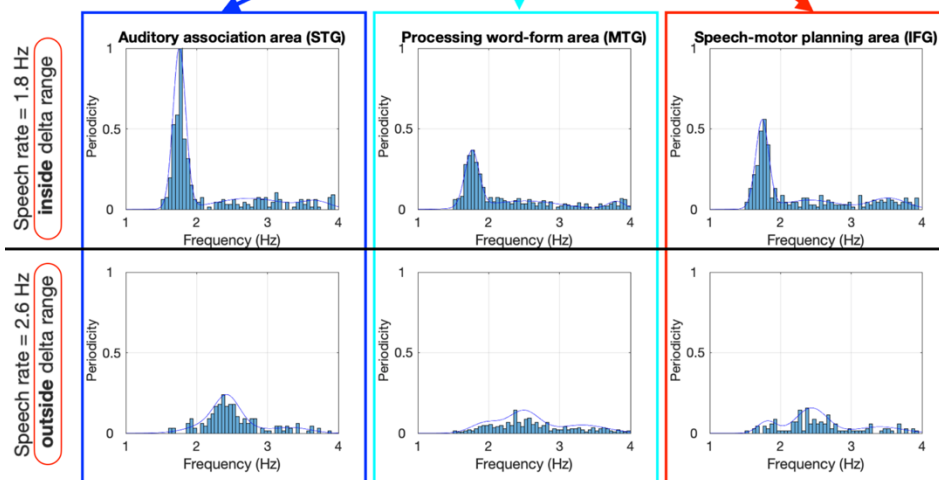


Figure 6c

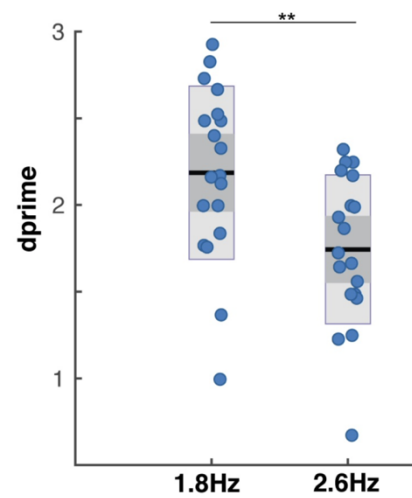


Figure 7a

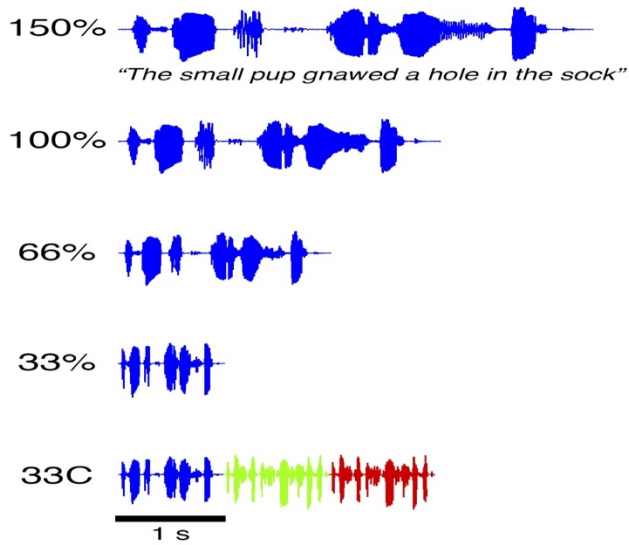


Figure 7b

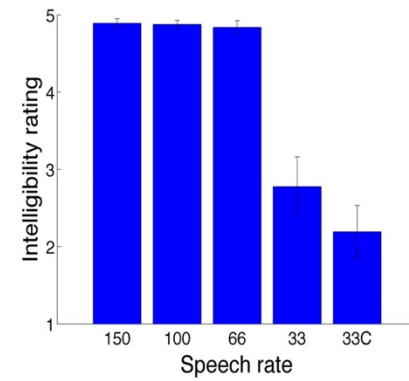


Figure 7c

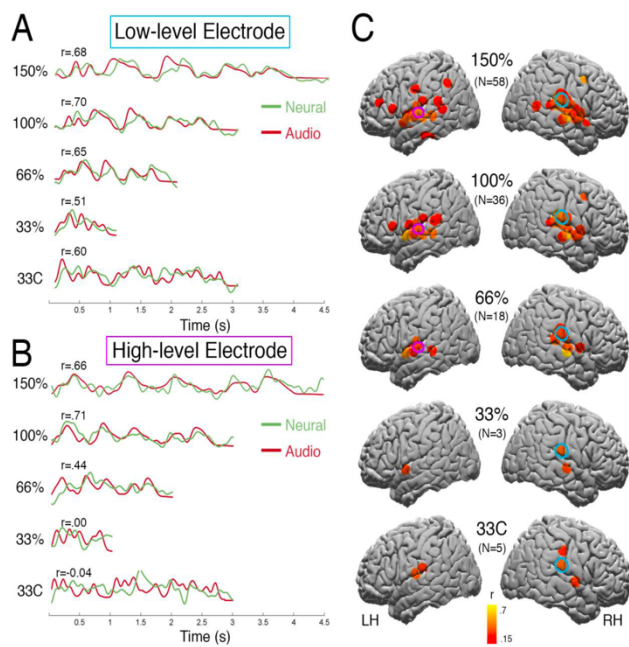


Figure 7d

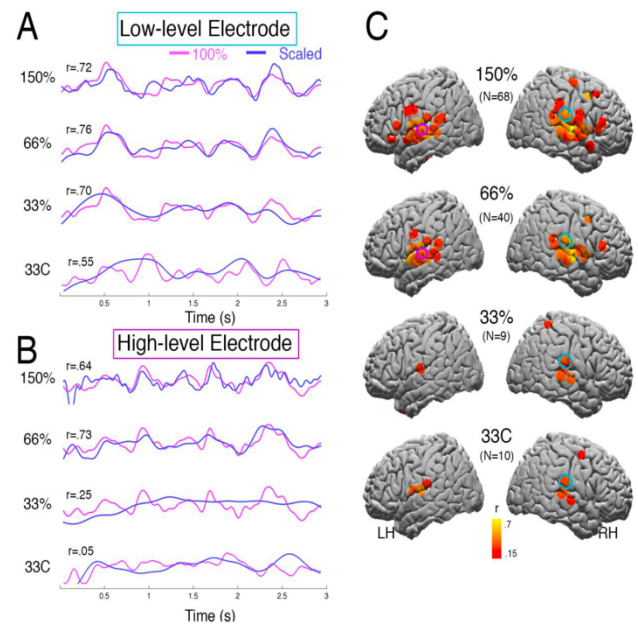


Figure 8a

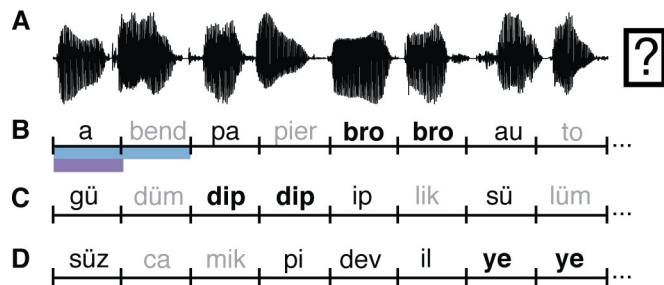


Figure 8b

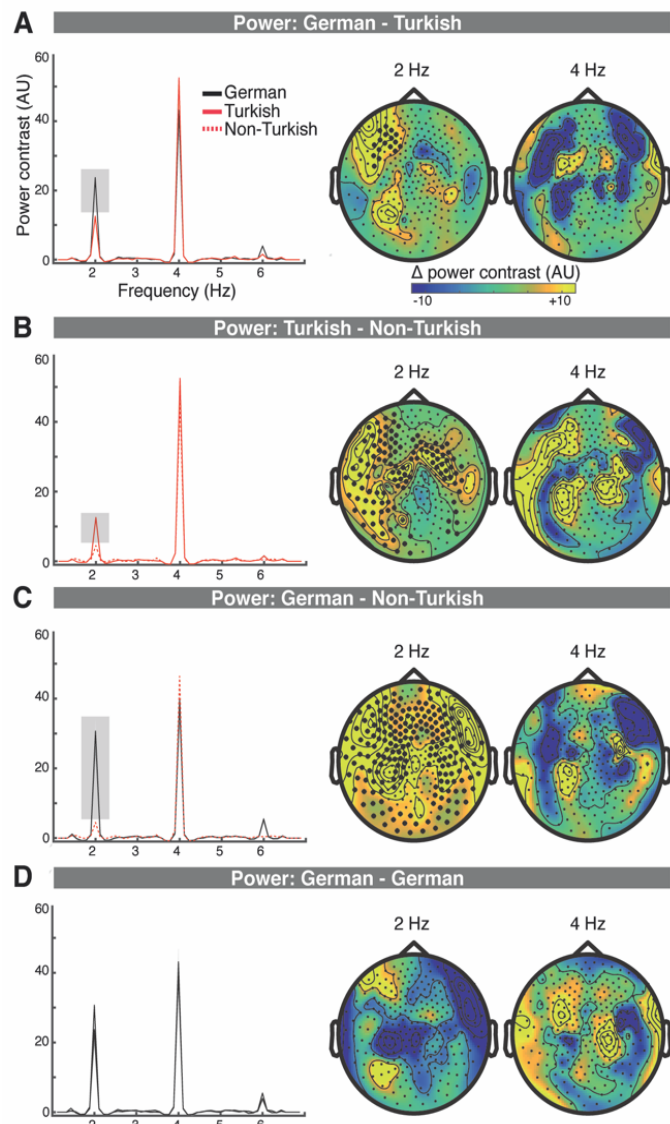


Figure 8c

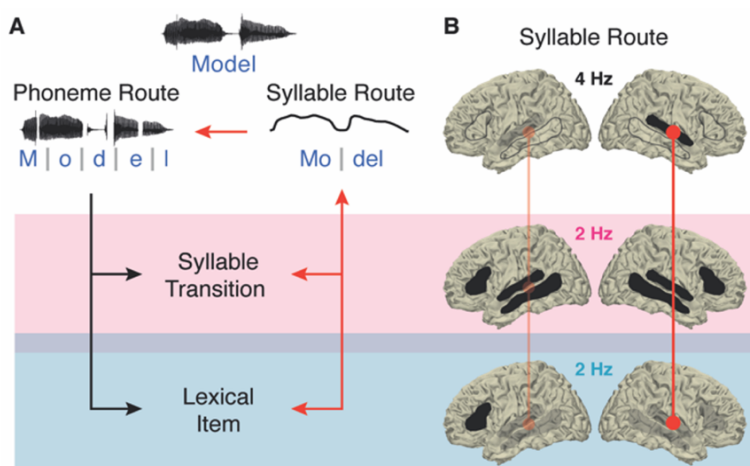


Figure 9a

(A) Typical sentence pair

1a. In his arms he held a bag of kibble.
1b. A drooling dog was sitting at home,
waiting to have his supper.

Prior context Local context Key word

(D) Prior context

P(dog | kibble) strong
P(bear | keys) weak

(E) Syntactic position

a
b
early late

(B) Target words

Defined by category (e.g., 🐕, 🐶, 🐾),
number of syllables and word frequency

(C) Local context

2-gram(dog drooling)/1-gram(drooling)
= 5/456 ≈ 0.1 > 0 = strong context
2-gram(bear lounging)/1-gram(lounging)
= 0/722 = 0 = weak context

(F) Story structure

64 📖 X 8 stories 📖 = 1.024 sentences 📖
Per 📖 8 X (2 X 2 X 2) conditions

Figure 9b

Sentence pair 1, story 1, male speaker

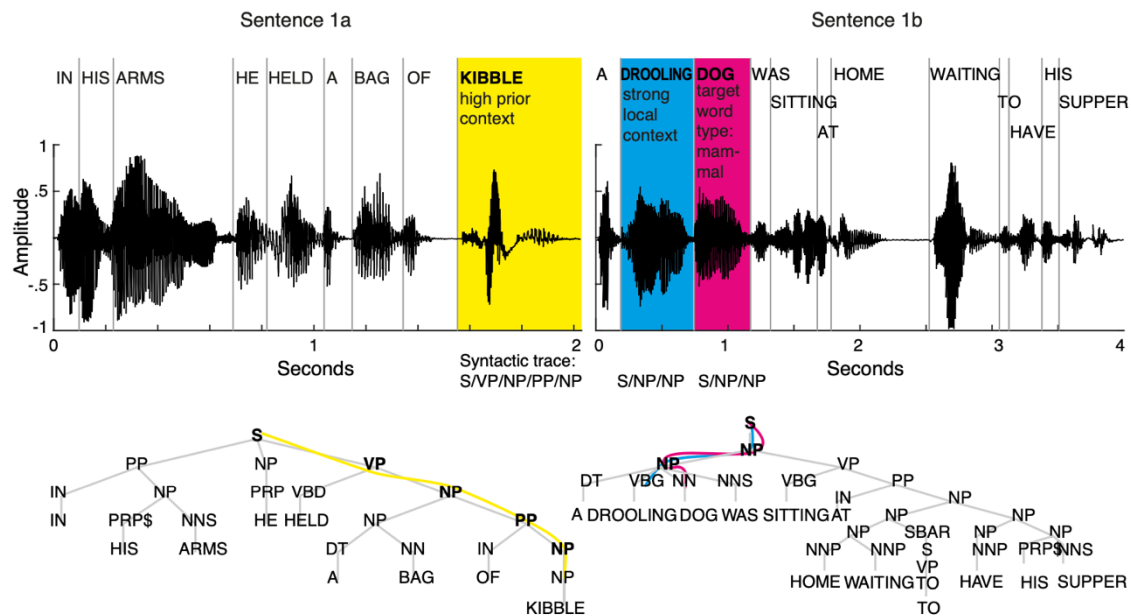


Figure 10

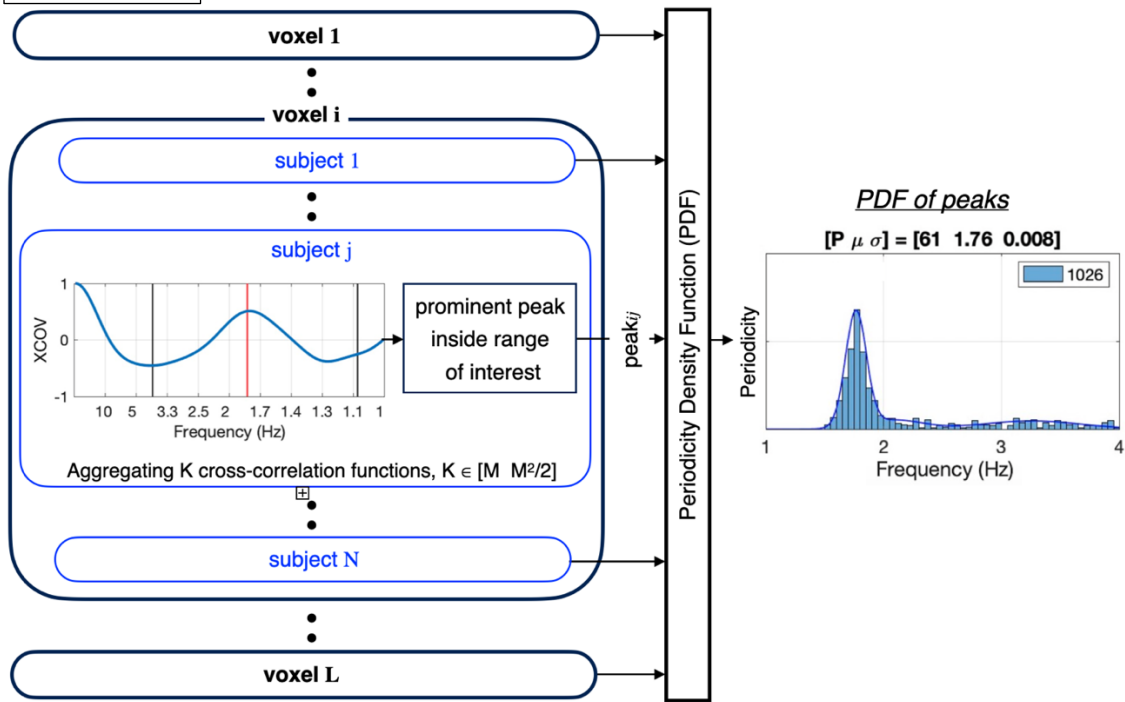


Figure 11

