


Acoustic-driven delta rhythms as prosodic markers

Oded Ghitza


To cite this article: Oded Ghitza (2016): Acoustic-driven delta rhythms as prosodic markers, Language, Cognition and Neuroscience, DOI: [10.1080/23273798.2016.1232419](https://doi.org/10.1080/23273798.2016.1232419)

To link to this article: <http://dx.doi.org/10.1080/23273798.2016.1232419>

 View supplementary material 

 Published online: 11 Oct 2016.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

Acoustic-driven delta rhythms as prosodic markers

Oded Ghitza

Department of Biomedical Engineering & Hearing Research Center, Boston University, Boston, MA, USA

ABSTRACT

Oscillation-based models of speech perception postulate a cortical computation principle by which decoding is performed within a time-varying window structure, synchronised with the input on multiple time scales. The windows are generated by a segmentation process, implemented by a cascade of oscillators. This paper tests the hypothesis that prosodic segmentation is driven by a “flexible” (in contrast to autonomous, “rigid”) oscillator in the delta range (0.5–3 Hz) by tracking prosodic rhythms, such that intelligibility is impaired when the ability of this oscillator to synchronise to these rhythms is impaired. In setting phrasal boundaries, both bottom-up acoustic-driven and top-down context-invoked processes interact in a manner that is difficult to decompose. The present experiments used context-free random-digit strings in order to focus exclusively on bottom-up processes. Two experiments are reported. Listeners performed a target identification task, listening to stimuli with prescribed chunking patterns (Experiment I) or chunking rates (Experiment II), followed by a target. 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. Importantly, performance remains high as long as the chunking rate is within the frequency range of neuronal delta, but sharply deteriorates for higher rates. This data provides psychophysical evidence for the role of acoustic-driven segmentation, with flexible delta oscillations at the core, in digit retrieval.

ARTICLE HISTORY

Received 16 January 2016
Accepted 19 August 2016

KEYWORDS

Prosodic segmentation; delta rhythms; synchronisation; chunking; word retrieval

1. Introduction


In written text, spaces and punctuation rules are used to divide words and define phrase boundaries. In contrast, naturally spoken language is a stream of connected sounds. Embedded in the acoustic stream is information analogous to spaces and punctuation rules (e.g. intonation, stress, pauses), termed “accentuation”, that is used by the listener to mark the boundaries of speech fragments associated with linguistic units. The marking is obtained by segmentation – a process by which the input signal is partitioned into temporal segments that are ultimately linked to a variety of linguistic levels of abstraction, ranging from phonetic segments to syllables to words and, ultimately, prosodic phrases. The segmentation process works on intervals associated with syllables (50–250 ms), termed here *Syllabic Segmentation*, and on the phrasal level (0.5–2 s), termed *Prosodic Segmentation*. Only after the signal has been segmented can *effective* decoding proceed. If the signal is incorrectly segmented, it is more difficult to form a match with internal linguistic patterns associated with syllables, words and phrases.

Before proceeding further a note on terminology is in order. In the context of perceptual prosodic segmentation, the term chunking is often used. Unfortunately,

this term is also used in the context of speech synthesis, to describe how acoustic chunks are produced. To avoid ambiguity, we shall use chunking for synthesis and segmentation for perception. Hence, *Chunking* refers to the process, executed by machine, of grouping acoustic items that are short in duration (e.g. phones, syllables, words) into one acoustic item (a chunk) – an operation that has nothing to do with perception. *Segmentation* is a cortical operation that results in an internal, temporal partitioning of the acoustic stream. Following this terminology, chunking results in acoustic *chunks* while segmentation guides a decoding process that results in *chunk-objects*. (See Glossary table, [Table 1](#).)

Turning to prosodic segmentation, this cortical process arises from the need to buffer the linguistic units carried by the acoustic stream into short-term memory (STM, e.g. Baddeley, 2010), a storage limited in capacity (measured in number of items per second). This limit implies that in order to maintain high performance, the incoming stream must be segmented into chunk-objects so that the rate of chunk-objects will not overwhelm the buffer store capacity. The prosodic segmentation process only marks the boundaries of temporal chunks that are likely candidates for a match with

CONTACT Oded Ghitza  oghitza@bu.edu

 Supplemental data for this article can be accessed at <http://dx.doi.org/10.1080/23273798.2016.1232419>.

© 2016 Informa UK Limited, trading as Taylor & Francis Group

Table 1. Glossary table.

Chunking	A process, executed by machine, of grouping acoustic items that are short in duration (e.g. phones, syllables, words) into one acoustic item (a chunk); an operation that has nothing to do with perception.
Segmentation	A bottom-up cortical operation, which sets a time-varying window structure synchronised to the input that results in an internal, temporal partitioning of the acoustic stream.
Parsing	A top-down cortical operation, which refers to the exhaustive division of the incoming speech signal into linguistic constituents using their syntactic roles (as part of the decoding process).
Chunk	An <i>acoustic</i> speech fragment generated by chunking.
Chunk-object	An <i>internal representation</i> of a chunk, generated by the decoding process which, in turn, is guided by segmentation and parsing.
Flexible oscillators	Capable of tracking syllabic irregularities – for example, a stress syllable followed by a non-stressed syllable (the theta range) – or slowly varying phrase irregularities (the delta range). In contrast to autonomous, “rigid” oscillators.
Acoustic-driven oscillators	Driven by sensory-input rhythms.
Context-driven oscillators	Driven by temporal regularities in past linguistic content.
Root strings	100 <i>text</i> strings – 10-digits long each – generated at random to form the root vocabulary of digit strings. Each experimental stimulus in this study is originated in one of the root strings. See Section 3.2.
Core stimuli	A vocabulary of “Lego” <i>acoustic</i> speech segments, from which the stimuli presented to the listeners are synthesised (by concatenation). See Section 3.2.
Chunking pattern	Digit strings are grouped into chunks – a chunk being multi-digit – with a prescribed chunking pattern. The sequence 3762895069 can be chunked into the regular chunking pattern [37 62 89 50 69], or into the irregular pattern [376 289 50 69]. See Section 4.1.1.
Chunking rate	The number of chunks per second, in Hz. See a rigorous definition in Section 4.2.1.
Prosody mode	For a prescribed root string and a prescribed chunking pattern two 10-digit long stimuli are synthesised, one for each of two prosody modes, <i>Gapped</i> and <i>Accentuated</i> . See a rigorous definition in Section 4.1.1.
Accentuation	Information embedded in the acoustic stream (e.g. intonation, stress, pauses) – analogous to spaces and punctuation marks in text – that is used by the listener to mark the boundaries of speech fragments associated with linguistic units.

what is stored in long-term memory (LTM). The linguistic identity of the chunk – the chunk-object – is obtained by decoding processes operating at phonetic, syllabic, lexical and phrasal levels. Pointing to yet another unfortunate terminology, note that this chunk-object – the internal representation of a chunk, stored as an item in STM – is different from chunk-objects obtained by “recording”, a fundamental concept introduced by Miller (1956), according to which a sequence of STM items can be learned if they are *contextually* recorded into chunk-objects before stored in LTM. Recording is beyond the scope of this study.

Miller was also concerned with identifying the decision units in perception of speech. In discussing the decision-making process in ordinary conversation he wrote (Miller, 1962):

The assumption ... is that people have available a relatively slow, single-channel mechanism for making decisions, so that it is necessary to store some of the input information and to process it as a unit. Decisions, therefore, occur at discrete points in time and serve to mark the boundaries for the units involved.

He then suggested:

Perhaps we make about one decision per second in ordinary listening. If we accept this as a rough estimate, it is suggested that the phrase – usually about two or three words at a time – is probably the natural decision unit for speech.

According to Miller, speech understanding is the result of delayed decisions over some “decision unit” that is multi-item in duration (about 1-sec long).

Interestingly, the duration of the decision unit suggested by Miller (1962) – about 1 second – coincides with Pickett and Pollack’s (1963) assertion that in read passages, and in ordinary conversation, a window of at least 1 second is required to reliably decode words, irrespective of the number of words presented. What makes a 1-sec-long window so special?

It is a commonplace observation that, in spoken word retrieval tasks, recall is improved when the material to be remembered is in some way organised or grouped in chunks. (This is why telephone numbers are typically grouped into 2, 3 digit chunks. Interestingly, chunking patterns vary across countries.) For example, Ryan (1969) studied several means of chunking – temporal, rhythmic, spatial – and reported an improvement in recall with grouping. The extent of this improvement depends on the size of the groups: when listening to digit strings, grouping into chunks of about three items each is most helpful (e.g. Chen & Cowan, 2005; Gilbert, Boucher, & Jemel, 2014, 2015; Maybery, Permentier, & Jones, 2002; Reeves, Schauder, & Morris, 2000; Ryan, 1969; Wickelgren, 1964).

And recently, neuroimaging evidence for neuronal activity associated with prosodic segmentation of spoken material has been found (e.g. Boucher, 2006; Buiatti, Peña, & Dehaene-Lambertz, 2009; Gilbert et al., 2014, 2015). EEG data showed that temporal groupings affect amplitude changes in N400 and P300 waves in a way that coincides with chunk-by-chunk segmentation (Gilbert et al., 2014). In a subsequent study they demonstrated that Positive Shifts in evoked potentials – which reflect neural responses to acoustic cues

arching over a group – are evoked by chunks marked by lengthening of final elements, on a chunk-by-chunk basis (Gilbert et al., 2015). Interestingly, a preferential response is recorded for chunks that are 3 items long, in accordance with the behavioural data noted above. What is the underlying cortical computation principle at the origin of superior performance for a grouping of about three items – roughly 1-sec long in duration?

Segmentation, prosodic and syllabic, is associated with distinct properties of the auditory response to the acoustic signal. The temporal fluctuations of the cochlear critical-band envelopes span two distinct time scales, one associated with slow modulations (i.e. < 3 Hz) and pertains to the prosodic stress patterns, and the other associated with faster modulations (3–20 Hz) and pertains to syllabic patterns. The faster scale conveys acoustic features important for decoding individual phonetic segments within a syllable. The slower scale conveys information on accentuations (e.g. intonation, stress, pauses) arching over linguistically related syllables and words, grouped into higher-level units associated with phrases.

As previously noted (e.g. Ghitza, 2011; Poeppel, 2003), there is a remarkable correspondence between average durations of speech units, on the one hand, and frequency bands of neuronal oscillations, on the other. Phonetic features (duration of 20–50 ms) are associated with beta (15–30 Hz) and gamma (>30 Hz) oscillations, syllables and words (mean duration of about 250 ms) with theta oscillations (3–9 Hz), and sequences of syllables and words embedded within a prosodic phrase (300–1500 ms) with delta oscillations (<3 Hz). Driven by this correspondence, it was proposed that neuronal oscillations play an important role in speech perception. A cortical computation principle was postulated, by which decoding is performed within a time-varying window structure, synchronised with the input on multiple time scales. The windows are generated by a segmentation process, implemented by a cascade of oscillators. In order to stay in sync with the quasi-regular rhythmicity of speech, a special class of oscillators is required – for example, the voltage-controlled-oscillator in a phase-lock-loop mechanism (e.g. Ahissar, Haidarliu, & Zacksenhouse, 1997; Viterbi, 1966; Zacksenhouse & Ahissar, 2006). Such oscillators are termed here *flexible oscillators*, in contrast to autonomous, rigid oscillators.

Syllabic segmentation pertains to speech fragments that are multi-phone in duration. Recent oscillation-based models of speech perception (e.g. Ahissar & Ahissar, 2005; Ding & Simon, 2009; Ghitza, 2011; Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012; Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015; Lakatos et al., 2005; Peelle & Davis, 2012; Poeppel, 2003) propose that syllabic

segmentation takes place in the pre-lexical layers, with processing time scales in the theta range. These models proved to be capable of explaining a range of counterintuitive psychophysical data (e.g. Ghitza, 2012, 2014; Ghitza & Greenberg, 2009) that are hard to explain by conventional models of speech perception. A computational model, TEMPO, which epitomises this computational principle is reviewed in Section 2.1.

Prosodic segmentation, which pertains to sequences of words associated with prosodic events, occur in higher cortical layers with processing time scales in the delta range. In setting phrasal boundaries, two distinct processes are at play, the contribution of each is hard to isolate: a bottom-up, acoustic-driven segmentation and a top-down, context-invoked parsing (*acoustic prosodic segmentation* and *contextual parsing* from here on). Cortical oscillations may be involved in the neuronal implementation of these processes. Hence, *acoustic-driven delta oscillations* may drive acoustic prosodic segmentation, while *context-invoked delta oscillations* may drive contextual parsing (e.g. Ding, Melloni, Zhang, Tian, & Poeppel, 2015). We distinguish between segmentation and parsing: the term segmentation refers to the function of setting a time-varying window – roughly 1-sec long – synchronised to the input, resulting in temporal partitioning of the acoustic stream. The term parsing refers to the exhaustive division of the incoming speech signal into refined candidate constituents using their syntactic roles (as part of the decoding process). In our view, acoustic prosodic segmentation and contextual parsing interact, with segmentation precedes parsing. The linguistic content – which drives the context-invoked delta – is provided by a decoding process guided by a flexible acoustic-driven delta, in the form of candidate linguistic constituents. Follows is a process guided by context-driven delta, which refines the division of the incoming speech signal.

The present study focuses exclusively on the cortical function that executes acoustic prosodic segmentation. In particular, we propose to generalise the cortical computation principle epitomised in the syllable level – that a (syllabic) segmentation process guides the syllable decoding process – to the chunk level, where a chunk is a speech fragment that is multi-word in duration. We hypothesise 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? To separate the role of acoustic

segmentation from that of contextual parsing, a context-free random-digit strings were used (to eliminate contextual effects).

- (2) Does the chunking strategy – how elements are grouped into chunks – play a role in digit retrieval? (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?)
- (3) Are hidden prosody cues – for example, 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, an adapted Sternberg task was used (1966). Listeners heard 10-digit utterances with different chunking patterns and prosody modes, followed by 2- or 3-digit-long targets, and were asked to indicate whether or not the target was part of the preceded utterance. The task is suitable for probing the interaction between segmentation and decoding in a memory retrieval task. This is so because a successful yes/no decision depends on how accurately the digit chunks are remembered, which in turn depends upon how accurately they are decoded, which depends on their correct segmentation.

Experiment I (Section 4.1) addresses 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. As we shall see, 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 II (Section 4.2) addresses 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. As we shall see, 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.

The remainder of the paper is organised as follows. The oscillation-based model is outlined in Section 2. Section 3 describes the experimental design, the core of the speech corpus,¹ the experimental paradigm and the data analysis methodology. The stimuli preparation and the results are described in Section 4. Finally, the interpretation of the data is discussed in Section 5 through the prism of oscillation-based models.

2. Segmentation with nested oscillations

We are concerned with the cortical function that executes segmentation of everyday speech (i.e. speech uttered in a continuous, natural way). With this focus in mind, the remainder of the paper adheres to a particular partitioning of the auditory system, driven by *function*:

Definition: The *auditory channel* includes all pre-lexical layers, with acoustic waveforms as input and syllable objects as output.

Corollary: The first layer of the *cortical receiver* is the lexical recognition circuitry (i.e. syllable objects as input and words as output).

Such a partitioning stems from the postulation that, when engaging in a spoken dialog, the smallest linguistically meaningful units are words (e.g. Cutler, 1994, 2012). According to this partition, the auditory channel includes all layers of the auditory periphery and the pre-lexical circuitry of TEMPO (black box in Figure 1). Syllabification takes place in the auditory channel, guided by a flexible oscillatory array with theta as the master, locked to the input rhythm. The sequence of output syllabic units (in the form of vowel – consonant-cluster – vowel, or VCV) is integrated into words and word sequences by a process that takes place in the cortical receiver, guided by delta oscillators.

One remark is worthy to note. Everyday speech is quasi-regular by nature in both the syllabic and the phrasal time scales.² In order to be able to stay in sync with this quasi-regular rhythmicity, the segmentation path of TEMPO is implemented by a special class of oscillators termed flexible oscillators. (Such oscillators are different in important respects from autonomous, rigid oscillators.) We argue that syllabic segmentation, accurate enough for a reliable decoding of pre-lexical units, can only be achieved by a flexible theta oscillator capable of tracking the syllabic irregularities (e.g. a stress syllable followed by a non-stressed syllable). Similarly, prosodic segmentation, accurate enough for a reliable decoding of phrases, can be achieved by a flexible delta oscillator capable of tracking slowly varying phrase irregularities.

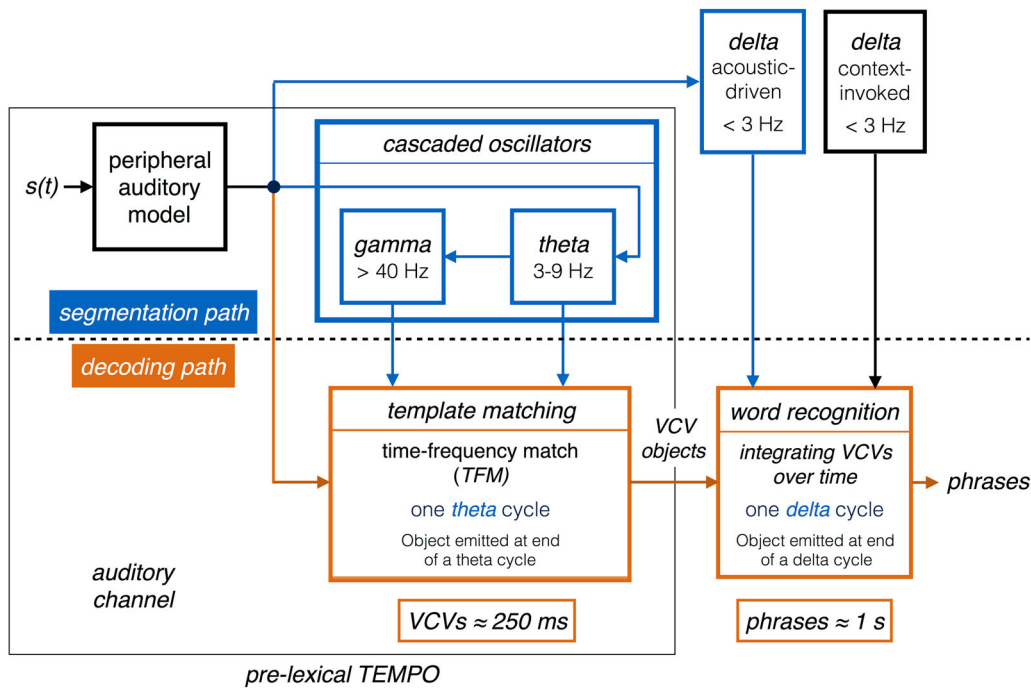


Figure 1. A block diagram of the TEMPO model. It comprises a decoding and a segmentation paths that process the sensory stream generated by a model of the auditory periphery. The decoding process (in orange) links chunks of sensory input of different durations with stored linguistic memory patterns, and it conforms to conventional models of speech perception. The segmentation path (in blue) generates a hierarchical window structure synchronised with the input, implemented by an array of cascaded oscillators locked to the input rhythm. The oscillators are assumed to be flexible, capable of tracking the slowly varying input rhythm. The instantaneous frequencies and relative phases of the oscillations determine the location and duration of the temporal windows that control the decoding process. The theta oscillator guides the decoding of pre-lexical VCV objects, and the sequence of VCV objects is integrated to form a delta-cycle-long phrases. The decoding process in pre-lexical TEMPO is performed during the theta cycle and a VCV object is emitted at the end of that cycle. The decoding at the phrase level is by integrating the VCV objects over the delta cycle, and a phrase is emitted at the end of that cycle. The segmentation path plays a crucial role in explaining the data by Ghitza and Greenberg (2009) and Ghitza (2014).

2.1. Syllabic segmentation steered by flexible theta

The black box within Figure 1 depicts the pre-lexical TEMPO, an oscillation-based model of the auditory channel (Ghitza, 2011). Conventional models of speech perception (up through the word level) assume a decoding of the acoustic signal by linking phonetic, syllabic, lexical and phrasal tokens in the auditory input with stored memory patterns (e.g. Luce & McLennan, 2005; Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; Stevens, 2005). These models have been shown to be incomplete. For example, they have a difficulty explaining the intricate pattern of human performance as a function of speech-speed and repackaging.³ Such data can be accounted for by TEMPO, a model which epitomises recently proposed oscillation-based models of speech perception (e.g. Ahissar & Ahissar, 2005; Ding & Simon, 2009; Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012; Hyafil et al., 2015; Lakatos et al., 2005; Peelle & Davis, 2012; Poeppel, 2003). The cortical computation principle at the core of TEMPO is that the speech

decoding process is performed within a time-varying, hierarchical window structure synchronised with the input, on multiple time scales. The window structure is generated by a segmentation path, implemented by a cascade of flexible oscillations with theta as “master”, capable of tracking the input pseudo-rhythm.⁴ A successful tracking can only be maintained if the input rhythm is within the theta frequency band. The frequencies and intra-phase configuration of the oscillators in the array determine the segmentation process. At the end of each theta cycle TEMPO outputs a VCV object, termed a theta-syllable.⁵

TEMPO is 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. Doelling, Arnal, Ghitza, & Poeppel, 2014; Ghitza, 2012, 2014; Ghitza & Greenberg, 2009). The key properties that enable such accountability are: (i) the capability of the theta oscillator – and hence the entire array – to track and stay locked to the input syllabic rhythm, and (ii) the cascaded nature of the oscillators within the array.

The tracking capability of the array maintains a match between the amount of information in the input stream (in terms of the number of syllables per unit time) and the capacity of the auditory channel (in terms of a reliable information transfer of VCV objects per unit time).⁶ Intelligibility remains high as long as theta is in sync with the input (as is the case for moderate speech speeds) and it sharply deteriorates once theta is out of sync (when the input syllabic rate is outside the theta frequency range).

2.2. Acoustic prosodic segmentation steered by flexible delta

In the present study, the cortical computation principle in play at the auditory channel is generalised to the cortical receiver. Brain's delta oscillations are hypothesised to be linked to chunks – multi-word in duration, analogous to the way neuronal theta oscillations are linked to VCV segments – multi-phone in duration. As shown in Figure 1, the VCV objects – the auditory channel output – are integrated to form words and phrases, guided by a prosodic segmentation process. The integration takes place during a temporal window that is one delta-cycle long. Intelligibility remains high as long as the delta cycles are aligned with chunks; this is so as long as the flexible delta oscillator is in sync with the chunk rhythm.

Two remarks are noteworthy. First, in analogy to the decoding process in pre-lexical TEMPO – where decoding is performed during the theta cycle and a VCV object is emitted at the end of that cycle – the decoding at the phrase level is by integrating the VCV objects over the delta cycle; a phrase is emitted at the end of that cycle. Second, given that decoding at the cortical receiver utilises information carried by linguistic structure, a contextual parsing driven by temporal regularities in past linguistic content – captured by a context-invoked delta – is also envisioned. Examining the possible role of context-invoked delta in contextual parsing, and the interaction between acoustic segmentation and contextual parsing, are beyond the scope of this study.

3. Methods

3.1. Experimental design overview

The hypothesised role of acoustic delta segmentation in decoding speech should ultimately be validated with experiments using continuous speech without linguistic constraints. Using such material, however, will result in contaminated data because of the difficulty to distinguish between bottom-up, acoustic segmentation and top-down, contextual parsing. Therefore, the

present experiments used context-free random-digit strings – 10 digits long – in order to focus exclusively on bottom-up processes. The digit strings to be presented were grouped into chunks – a chunk being multi-digit in duration.

Two experiments were conducted. In Experiment I, the digit strings were chunked to a prescribed chunking pattern. For example, the sequence 3762895069 can be chunked into the regular chunking pattern [37 62 89 50 69], or into the irregular pattern [376 289 50 69]. Two chunking procedures were used, each characterised by a prosody mode – gapped or accentuated. In Experiment II, the chunking rate of a digit string was controlled, to be inside, or outside of the cortical delta frequency band (about 0.5–3 Hz). Note that these chunking operations only introduces a prescribed temporal structure, without any contextual advantage.

Error rate was measured using a retrieval task in the form of an adapted Sternberg target identification task (*target ID task* from here on): listeners 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 preceding utterance. Three target positions were considered: (i) *target inside* a chunk, (ii) *target split* between two successive chunks, and (iii) *no target* present in the 10-digit string. The task is suitable for probing the role of acoustic prosodic segmentation in a memory retrieval task. This is so because a successful yes/no decision depends on how accurately the digit chunks are remembered, which in turn depends upon how accurately they are decoded, which depends on their correct segmentation.

In Experiment I (Section 4.1), error rate was measured as a function target position (inside, split and none), with *chunking pattern* and *prosody mode* (gapped or accentuated) as the parameters, and while listening to stimuli in a normal speed (see Section 4.1.1 for the stimulus preparation details). A large error rate in the target-split condition, compared to that in the target-inside condition, will demonstrate the benefit of accurate acoustic prosodic segmentation (in the chunk level) in word retrieval. Such data, however, will not provide any evidence for the possible role of acoustic delta oscillations in prosodic segmentation.

In Experiment II (Section 4.2), error rate was measured with *chunking rate* the parameter. In particular, one of the target positions, the target-inside condition (where the target is always inside the chunk), provides an insight into the possible role of acoustic delta oscillations: a jump in error rate for chunking rates greater than the upper bound of the delta frequency-band (about 3 Hz) will support our hypothesis that a necessary condition for good performance is a successful synchronisation between the input and the delta oscillator.

3.2. Root strings and core stimuli

The 10-digit long stimuli were synthesised as follows. First, 100 text strings – 10-digits long each – were generated at random to form the root vocabulary of digit strings. These strings, termed *root strings*, are semantically unpredictable but of low perplexity (a vocabulary of 11 words, 0 to 9 and O). Using the AT&T Text-to-Speech System,⁷ with the female speaker Crystal, a bank of *core stimuli* was generated, populated with the following high quality, naturally accentuated stimuli:

- (1) One stimulus for each *single-digit* (0 to 9, and O).
- (2) One stimulus for each *doublet* of digits that exists in the entire set of root strings.
- (3) One stimulus for each *triplet*, *quartet*, and *quintet* of digits.

To synthesise a particular 10-digit stimulus, the sequence of core stimuli – defined by the prescribed chunking pattern – was concatenated, as described in Section 4.1.1.

3.3. Experimental stimuli

Experiment I and Experiment II were conducted separately. The signal processing used to generate the 10-digit long stimuli unique to each experiment are detailed in Sections 4.1.1 and 4.2.1. For every condition, 20 root strings (out of the total number of 100) were randomly selected. For each experiment, all stimuli–across all conditions–were scrambled, and the resulting pool of stimuli was divided into bundles, 65 stimuli per bundle. The 65 stimuli (in a bundle) were concatenated in the following sequence, to be presented to the listener: [alert tone] [1-sec long silent gap] [digit string] [1-sec long silent gap] [target] [3-sec long silent gap] [alert tone] [1-sec long silent gap] [digit string] ..., resulting in one concatenated audio clip, about 8-minutes long. Same bundles were presented to all subjects (i.e. per condition, all subjects heard the same 20 stimuli).

3.4. Subjects

All listeners, 18 in number, were young adults (college students), educated in the USA (English as first language), with normal hearing (screened for normal threshold audiograms). Nine listeners (five female and four male students) participated in Experiment I, and nine (five female and four male students) in Experiment II. The responses in each experiment were reasonably consistent with each other, hence no further recruitment was needed.

A participant provided hers/his written informed consent to participate in this study. The human-subjects protocol for this study (including the informed consent document) was approved by the Institutional Review Board of Boston University.

3.5. Experimental paradigm

Subjects performed the experiments in an isolated office environment (no other occupants) using headphones. The sound pressure was adjusted by the subject to a comfort level and remained unchanged throughout the experiment. Stimuli were presented diotically. In a session, subjects heard 7 audio clips – a clip being the sequence of stimuli concatenated as described in Section 3.3 (approximately 8 minutes long). They were instructed to listen to each audio clip uninterrupted and to type into a text file a 1 or a 0 during the 3-sec long gap following a target, indicating whether or not the target was part of the preceded digit string. Subjects were not informed about the various chunking conditions and no feedback was provided. A subject participated in four sessions in completion.

3.6. Data analysis

Depending on the clarity of the emerging error patterns, data is presented either as the error rate of the raw data, or as the outcome of a hierarchical logistic regression used to model the data.

When error patterns of a phenomenon considered are clear and demand no further quantification, accuracy is presented as the mean and standard deviation of the error rate across subjects, plotted as bar charts. Each bar shows the error rate and the standard deviation⁸ calculated over 20 responses.

For other conditions, with less obvious data, a detailed evaluation is performed to assess accuracy in terms of predictions of a hierarchical logistic regression used to model the data. The model is derived as follows. Per stimulus, target identification is defined as x_i , with $x_i=1$ when identification is correct and 0 otherwise. Per experiment, the data comprises 9 subjects, each of which was tested under N conditions, $\psi \in \{1, 2, \dots, N\}$, with 20 sentences heard under each condition. (For example, in the upper left panel of Figure 6, per chunking pattern condition (333, 22222, etc.), ψ is the chunking rate with $N=4$, that is, $\psi \in \{[2-2.5], [2.5-3], [3-3.5], [3.5-4]\}$ Hz.) A hierarchical logistic regression was used to model the data, capturing the effect of each subject and each condition ψ on target identification. This approach is conceptually similar to a classical ANOVA comparison (Gelman, 2005): (a) inferences for all means

and variances are performed under a model with a separate batch of effects for each row of the ANOVA table; (b) the model automatically gives the correct comparisons even in complex scenarios; and (c) this is a preferred approach when dealing with small sample size, as is the case here with only 9 subjects.

The model provides estimates for the average accuracy at each level of ψ . Instead of simply reporting standard errors for significance testing, this approach allows the flexibility of fully propagating the uncertainty inherent in all pieces of the model (Gelman & Hill, 2007). This is done through a simulation framework, where the model estimates are simulated 1000 times. We compute 95% credible intervals around the accuracy levels at each ψ – these are the Bayesian equivalent of confidence intervals, again accounting for the full uncertainty in the model.⁹

The results plotted are estimates of error rate with the 95% credible intervals, shown as a bar chart, calculated over 20 responses per bar. Visually, we emphasise the credible interval around the estimated error rate of ψ^* – the reference condition. The estimated error rates of the surrounding conditions are compared to the estimated error rate of the reference condition, and the credible intervals indicate whether the differences are statistically significant.

4. Stimuli and results

4.1. Experiment I: the role of acoustic prosodic segmentation

4.1.1. Stimulus preparation

Stimuli were generated for a number of conditions specified by two parameters, *chunking pattern* and *prosody mode*, defined below. Fifteen chunking patterns, listed in the abscissa of Figure 3, were used. The chunking patterns form regular and irregular grouping patterns, for example, 22222 (regular) or 424 (irregular).¹⁰ For any selected root string (out of the 100 10-digit root strings in the vocabulary, see Section 3.2), a prescribed chunking pattern defines the individual chunks. For example, for the root string 3762895069 and chunking pattern 442, the individual chunks are 3762, 8950, and 69. Once a root string and a chunking pattern were chosen, two 10-digit long stimuli were synthesised, one for each of two prosody modes, *Gapped* and *Accentuated*:

- (1) *Gapped mode*. For the 11111 chunking pattern,¹¹ the 10 single-digit core stimuli were concatenated with a 160-ms long silent gap inserted in between each digit stimulus. For the 3322 chunking pattern, for example (e.g. 678 903 21 56), each chunk was synthesised by grouping isolated single-digit core

stimuli as follows. For the first chunk, 678, the three single-digit core stimuli 6, 7, and 8 were concatenated, with a 10-ms long silent gap inserted in between them. The resulting 4 chunks were then concatenated, with a 160-ms long silent gap inserted in between the chunks. The average duration of the stimuli generated in this mode is 4.32 seconds, with a standard deviation of 0.46 seconds. The top panel of Figure 2 shows the resulting stimuli for root string 3323443215 synthesised in a Gapped mode, with a chunking patterns 3322. Note that a stimulus synthesised in the Gapped mode belongs to the “temporal grouping” condition of Ryan (1969).

- (2) *Accentuated mode*. Stimuli in this mode were generated by a procedure similar to the Gapped procedure, with two exceptions: (i) for a given chunking pattern, each chunk was the corresponding (accentuated) core stimulus (a doublet, a triplet, etc.), rather than a chunk created by concatenation. For example, for the 3322 chunking pattern above, the first chunk, 678, is the core triplet stimulus 678; and (ii) the accentuated chunks were separated by 10-ms long gaps, rather than 160-ms long gaps. Note the absence of the chunking pattern 11111 here (no accentuation for a single digit). The average duration of the stimuli generated in this mode is 3.51 seconds, with a standard deviation of 0.27 seconds. The bottom panel of Figure 2 shows the resulting stimuli (waveform and a Fourier spectrogram) for the root string 3323443215 synthesised in the Accentuated mode, with the chunking patterns 3322. A stimulus synthesised in this mode could be considered as belonging to Ryan (1969) “non-temporal grouping” condition since no explicit grouping have been applied. However, an implicit grouping exists, enforced by accentuation.

Two remarks are noteworthy. First, since the stimuli were synthesised in normal speed, the average chunking rate was well inside the delta frequency band (about 0.5–3 Hz) for all 15 chunking patterns and both prosody modes. Second, the reason for using two prosody modes in this experiment stemmed from the need to expand on Ryan’s distinction between temporal vs. non-temporal grouping (Ryan, 1969). Although an explicit inspection of a waveform synthesised in the Accentuated mode places it in the non-temporal grouping category, a question is raised about the possibility of an *implicit* temporal grouping via hidden prosody cues (i.e. accentuations arching over a chunk).

The experimental conditions for Experiment I are summarised in Table 2.

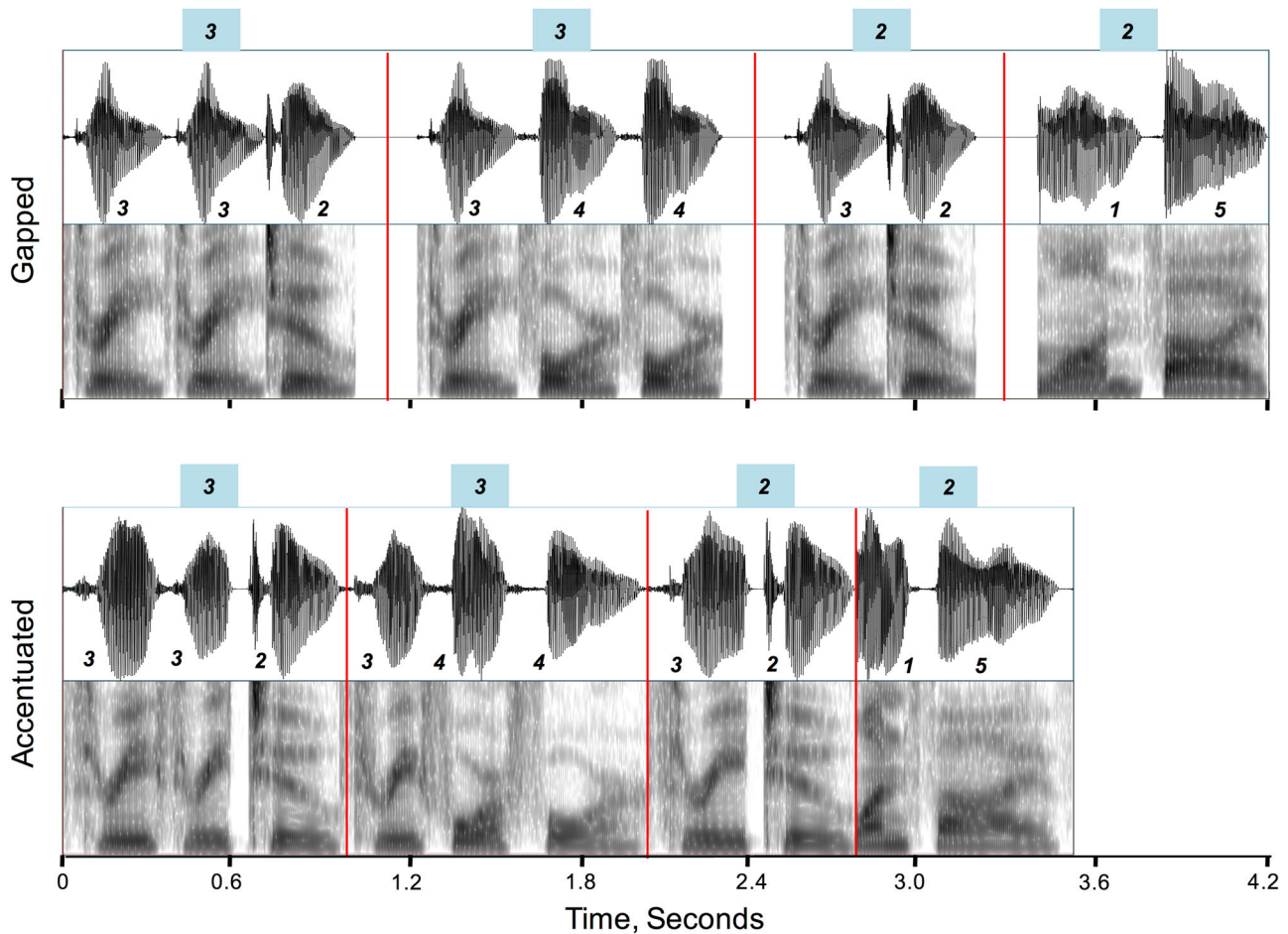


Figure 2. Stimuli used in Experiment I. Shown are waveforms and Fourier spectrograms for the root string 3323443215 with a chunking pattern 3322 (i.e. with the chunks 332 344 32 15), synthesised in a Gapped (top) and Accentuated (bottom) modes. The gap durations are 160-ms long for Gapped, and 10-ms long for Accentuated. Note that the chunk waveforms are uncompressed. (MP3 files are available for listening as Supplementary Materials.)

4.1.2. Data

Figure 3 shows error rate as a function of chunking pattern for the two prosody modes. Data are organised in three panels, one for each target position (inside, split and no target). The 2- and 3-digit target length conditions were combined. Each bar shows the error rate and the standard deviation calculated over 20 responses. For each panel, the mean and standard deviation across chunking patterns are shown on the left-hand side. Three observations are noteworthy:

- (1) A significant increase in error rate is observed for target split, compared to that for target inside. This is the case for all chunking patterns and both prosody modes.
- (2) No chunking pattern “stands-alone”, that is, no pattern provides a significant advantage in performing the task. This is the case for any target position.
- (3) A similar error pattern is observed for both prosody modes, Gapped and Accentuated.

Table 2. Summary of experimental conditions.

	Chunking pattern	Prosody mode	Chunking rate
Experiment I	<ul style="list-style-type: none"> • a parameter – 15 patterns (see abscissa of Figure 3) 	<ul style="list-style-type: none"> • a parameter – gapped – accentuated 	<ul style="list-style-type: none"> • not a parameter – 1–2 Hz – inside delta range
Experiment II	<ul style="list-style-type: none"> • a parameter – 2 regular patterns – 2 irregular patterns (see abscissa of Figure 5) 	<ul style="list-style-type: none"> • not a parameter – accentuated 	<ul style="list-style-type: none"> • a parameter – 2–4 Hz – 4 sub-bands, inside and outside delta range (see Figure 5 Legend)



Figure 3. Experiment I data. Error rate as a function of chunking pattern and prosody mode for the three target positions, target-inside a chunk (top panel), target-split between two successive chunks (middle panel), and no-target present (bottom panel). The 2- and 3-digit target length conditions are combined. The mean and standard deviation across chunking patterns are shown on the left-hand side of each panel. A significant increase in error rate is observed for target-split, compared to that for target-inside. This is the case for all chunking patterns, in both prosody modes. Also, no chunking pattern is preferred.

Since the difference between the target-split error rate and the target-inside error rate is large, and since there is no stand-alone pattern, no further statistical analysis was required to quantify these observations.

Our interpretation of these results will be discussed in Section 5.

4.2. Experiment II: the role of acoustic delta in segmentation

4.2.1. Stimulus preparation

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. For a particular root string (out of the 100 10-digit root strings in the vocabulary, see Section 3.2),

for each [chunking pattern]×[chunking rate] combination a 10-digit stimulus was synthesised using the Accentuated mode described in Section 4.1.1. Chunking rate was controlled by varying the silent gaps inserted 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 band. Figure 4 depicts the resulting stimuli for 20-ms (top) and 160-ms (bottom) gaps. Chunking rate (in Hz) is defined as the average of the inverse of chunk-duration of all chunks in the stimulus, where a chunk duration is the fragment between the mid points of two successive gaps (Figure 4, red markers). For example, the chunking rates of the two stimuli in Figure 4 are 4.2 Hz (top) and 2.8 Hz (bottom). To be able to generate stimuli with chunking rates greater than the upper bound of the delta frequency range

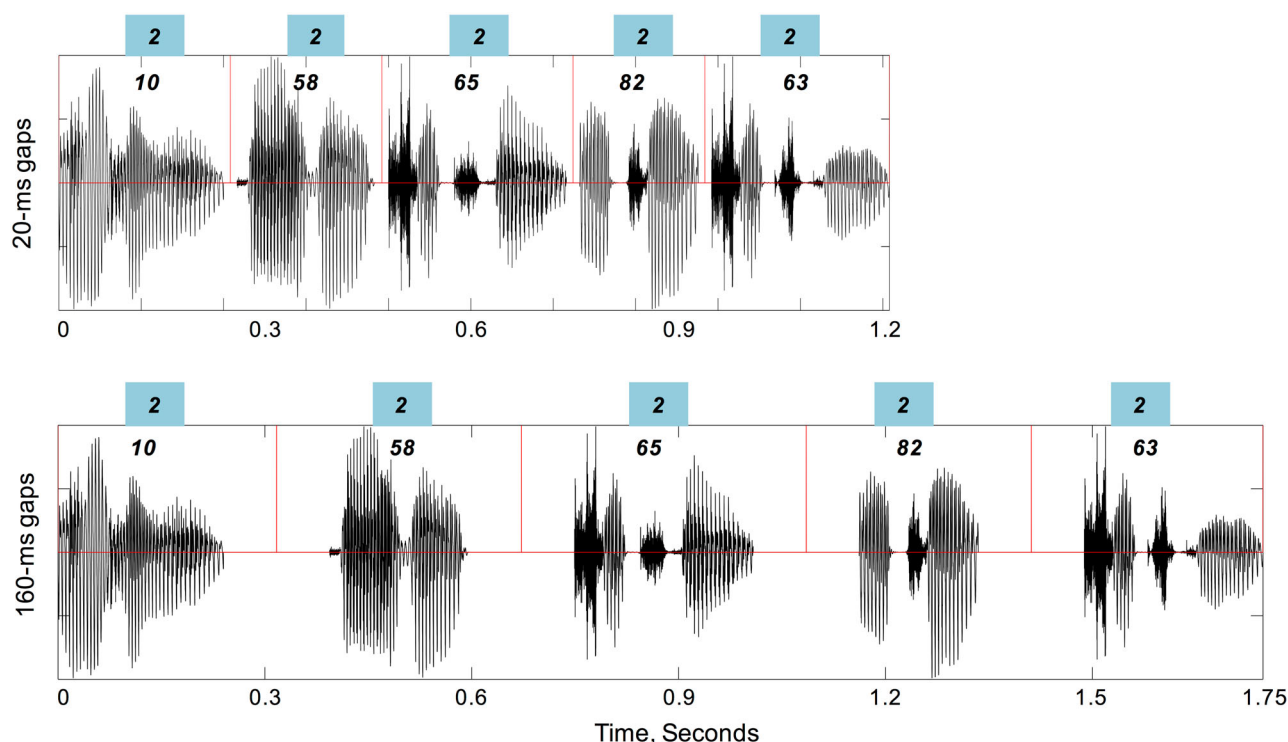


Figure 4. Stimuli used in Experiment II. Shown are waveforms – synthesised in the Accentuated mode – for the root string 1058658263, with the chunking pattern 22222 (i.e. with the chunks 10 58 65 82 63), with a 20-ms long silent gap, hence a chunking rate of 4.2 Hz (top), and a 160-ms long gap, hence a chunking rate of 2.8 Hz (bottom). Note that the waveforms within a chunk are time-compressed by a factor of 3 (MP3 files are available for listening as Supplementary Materials.)

(about 3 Hz), chunk (and target) waveforms were time compressed by a factor of 3 – just below auditory channel capacity (Ghitza, 2014).¹² Figure 4 shows the resulting stimuli for the root string 1058658263 synthesised in Accentuated mode and chunking pattern 22222, with a 20-ms long silent gap (top) and a 160-ms long gap (bottom). Note that the waveform of each chunk is time-compressed by a factor of 3.

The experimental conditions for Experiment II are summarised in Table 2.

4.2.2. Data

Figure 5 shows the error rate as a function of chunking pattern, for four bands of chunking rate (2–2.5, 2.5–3, 3–3.5 and 3.5–4 Hz), three target positions (inside, split and no target), and the two target lengths (2- and 3-digit-long). Each bar shows the error rate and the standard deviation calculated over 20 responses. For all [chunking pattern]×[target length] conditions, and for all chunking rates, errors in the target-split position are considerably larger than the errors in the target-inside position. No further statistical analysis was performed to quantify the significance of this error difference because of its large magnitude. Note that this observation concurs with a similar result observed in Experiment I (observation no. 1 in Section 4.1.2).

The error patterns for the target-inside position are not as obvious and required further quantification. For this purpose, the statistical analysis described in Section 3.6 was used. The data – shown in Figure 6 – are organised in a 3×2 matrix of panels, with target length as rows and chunking pattern as columns. In the left column, all chunking patterns are detailed. In the right column chunking patterns are collapsed into regular and irregular patterns. For each chunking pattern condition, estimates of error rate and the 95% credible intervals are shown as a bar chart, calculated over 20 responses per bar, with chunking rate as the parameter. The highest chunking rate condition – with the credible interval around it – is visually highlighted (gray horizontal strip). (Note that some bars are not valid: (i) all bars in [chunking pattern 22222]×[target length=3], because all chunks are shorter than the target, and (ii) the bar corresponding to the highest chunking rate (>3.5 Hz) in chunking pattern 333, because such stimuli can not be generated.)

In general, a consistent error pattern emerges across panels showing an increase in error rate with the increase of chunking rate, with a considerable jump in error rate at the highest chunking rate. The 95% credible intervals indicate that the differences in estimated error rates are statistically significant. The magnitude of the

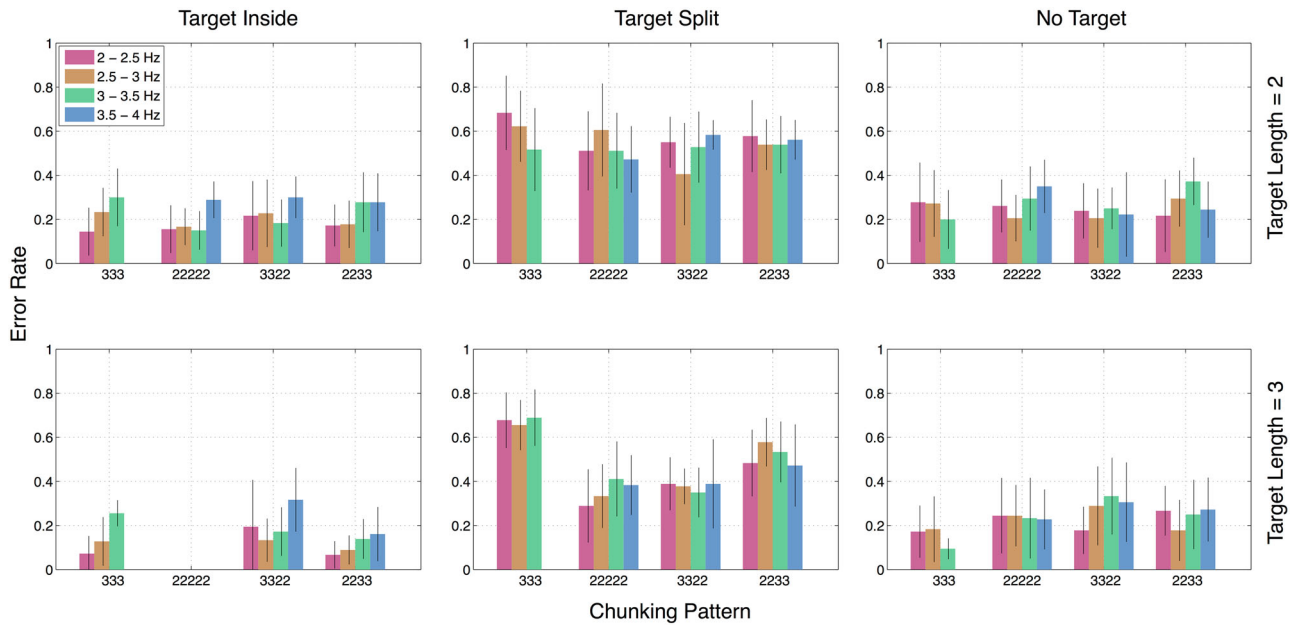


Figure 5. Experiment II data. Error rate as a function of chunking pattern for four bands of chunking rate (2–2.5, 2.5–3, 3–3.5 and 3.5–4 Hz), for the three target positions – target-inside a chunk (left panel), target-split between two successive chunks (middle panel), and no-target present (right panel) – and for 2- and 3-digit target lengths. A significant increase in error rate is observed for target-split, compared to that for target-inside. This is the case for all chunking patterns and chunking rates, in both target lengths. This error pattern concurs with the error pattern observed in Experiment I (Figure 3).

jump in error rate – and its statistical significance – depend on [chunking pattern]×[target length] condition. A few observations are noteworthy:

- (1) Bottom-left panel (pooled target lengths). A considerable jump is shown at the highest chunking rates, for all chunking patterns. In particular, note the jump magnitude for the regular patterns and for the irregular 3322 pattern – the pattern the (American) subjects are mostly used to.
- (2) Green bars (that is, rate = [3 – 3.5] Hz) for patterns 22222 and 333. When pooled together – to form the green bar for the regular pattern condition, the significant difference in error rate between these conditions is averaged out (top-right and bottom-right panels).

Our interpretation of these results will be discussed in Section 5.

5. Discussion

There is a body of work on the effects of chunking on recall, in various sensory modalities (e.g. visual, verbal). When recalling a sequence of words from a spoken utterance, listeners set phrasal boundaries to create groups so as to overcome capacity limitations in STM by partitioning the serial input into segments, and by linking the

segments to phrasal units. Two distinct processes are at play, the contribution of each is hard to isolate: a bottom-up acoustic segmentation and a top-down contextual parsing. The present study focuses exclusively on acoustic prosodic segmentation. In particular, we aim at providing psychophysical evidence for the role of a neuronal acoustic-driven delta oscillation in segmentation. In order to eliminate the effect of contextual parsing the present experiments used context-free random-digit strings. The task used – a target ID task – is suitable for testing the role of oscillations in segmentation (as reasoned in Section 3.1). In the remainder of this Section we interpret the results reported in Experiments I (Section 4.1.2) and II (Section 4.2.2), and hypothesise possible generalisations of the results to acoustic prosodic segmentation of unconstrained continuous speech.

5.1. The role of acoustic prosodic segmentation (Experiment I)

First, we note the significant increase in error rate for the target split condition, compared to the error rate in target inside. This result confirms the benefit in accurate bottom-up, acoustic prosodic segmentation (in the chunk level) in word retrieval: for a successful template matching – an operation at the core of the identification process (at the tail of a target-ID task) – a proper segmentation of the chunk is essential. This result may also be

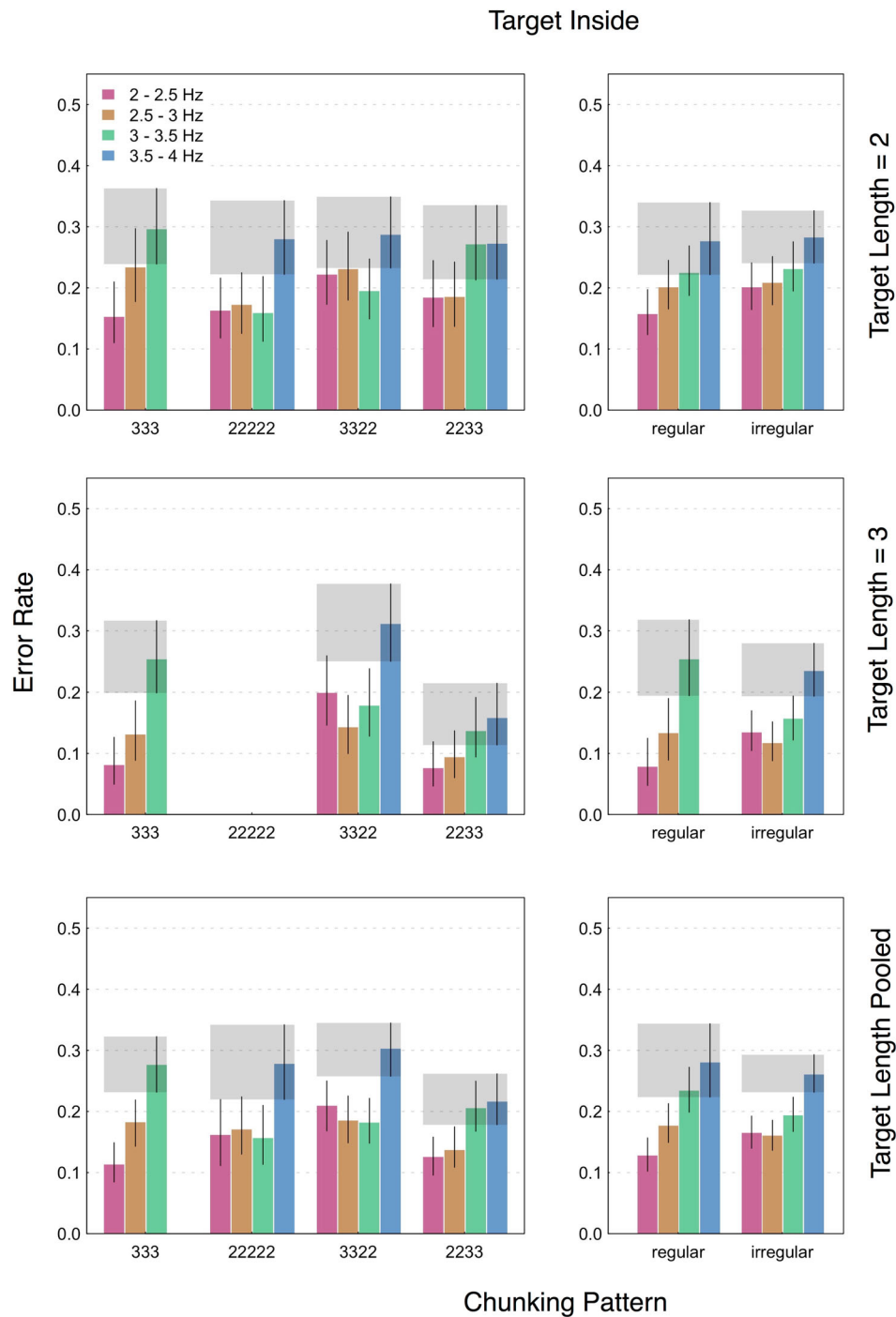


Figure 6. A quantified version of the target-inside condition in Figure 5, derived by the data analysis described in Section 3.6. (Note the different scaling of the ordinate compared to Figure 5.) In the right column chunking patterns are collapsed into regular and irregular patterns. For each chunking pattern condition, estimates of error rate and the 95% credible intervals are shown as a bar chart, with chunking rate as the parameter. The highest chunking rate condition – with the credible interval around it – is visually highlighted (gray horizontal strip). Across panels, error rate increases with the increase of chunking rate, with a considerable jump in error rate at the highest chunking rate. The 95% credible intervals indicate that the differences in estimated error rates are statistically significant. The magnitude of the jump – and its statistical significance – depend on the [chunking pattern]×[target length] condition. Note that some bars are missing – see reasoning in text (Section 4.2.2).

interpreted as an evidence that, in a digits retrieval task, a single-layer process is in play: chunks (and targets) are retrieved as whole objects, as opposed to an item-by-item retrieval process, as Miller (1962) posited.

Second, no stand-alone chunking pattern was observed. As long as chunking rate is within the cortical delta frequency band (about 0.5–3 Hz) – the case for all 15 chunking patterns tested here – a perception

tolerance to chunking pattern emerges. Notably, our data do not show a meaningful advantage for grouping in threes, as shown by others (e.g. Chen & Cowan, 2005; Gilbert et al., 2014, 2015; Maybery et al., 2002; Reeves et al., 2000; Ryan, 1969; Wickelgren, 1964). An interesting corollary to this observation concerns the manner by which telephone numbers are chunked in different countries. In all languages, at normal speech-speed the chunking rate is inside the cortical delta band. If we assume same delta band across gender and race (indeed species; e.g. Buzsaki, Logothetis, & Singer, 2013) this result suggests that chunking strategy is of cultural consequence rather than the result of a need to match a cortical constraint. Interestingly, for our American listeners – educated in the USA with English as their first language – the pattern 3322 (the one used by Americans) was not advantageous.

Third, a similar error pattern for both prosody modes, Gapped and Accentuated, was observed. This result suggests that hidden prosody cues – accentuations arching over a chunk – result in grouping with a benefit equivalent to the benefit gained by explicit temporal grouping (i.e. by inserting gaps). Robust acoustic correlates to the accentuation arch, however, are yet to be discovered.

Finally, performance in the no-target condition is a measure of the degree of guessing by the subject. The results in the no-target condition confirm that guessing is uniform with chunking pattern.

5.2. The role of acoustic delta in segmentation (Experiment II)

Similar to what was observed in Experiment I, a significant increase in error rate is registered for target split, compared to the error rate in target inside. This data reinforces the corresponding interpretation in Section 5.1, that is, that a proper segmentation of a chunk is essential for digit retrieval.

Importantly, Experiment II also provides behavioural evidence for the role of acoustic delta in segmentation. For target inside, error rate moderately increases with the increase in chunking rate, with a significant jump for chunking rates greater than the upper bound of the delta frequency band (about 3 Hz). This data supports our hypothesis that a successful segmentation is possible only if the chunking rate complies with the range of delta frequencies. From time compression studies (e.g. Ghitza, 2011; Vagharchakian, Dehaene-Lambertz, Pallier, & Dehaene, 2012) we know that a reliable retrieval requires a sufficient decoding time. Extra decoding time, if needed, can be provided by the insertion of silent gaps. As was shown before (e.g. Ghitza, 2014;

Ghitza & Greenberg, 2009), at the syllable level, the decoding time necessary for a reliable decoding is determined by theta. We postulate that the gap duration that ensures reliable retrieval of chunks is determined by delta: the duration of “the chunking cycle” – the concatenation of the chunk and the following gap – must be greater than a threshold cycle duration, equals the inverse of the upper bound of the delta frequency band (~330 ms).

Three remarks are noteworthy. First, performance is a reflection of the role of prosodic segmentation, in isolation from syllabic segmentation. This is so because the acoustics inside a chunk is unchanged for all chunking rates \Rightarrow the syllabic rate inside a chunk is the same for all chunking rates \Rightarrow the predicted cortical theta, locked to the input syllabic rate, is the same for all chunking rates. Second, performance is a reflection of the role of acoustic delta, in isolation from context-invoked delta, because the study was confined to the task of digit retrieval when listening to digit strings – a material with no context. Third, a question may be raised about performance when chunking rate is slower than the lower bound of the delta range (~0.5 Hz). The reason to skip the effects of time dilation stemmed from the fact that memory decay time – about 2 sec long (e.g. Cowan, 1984) – roughly coincides with the lower bound of the cycle duration in delta band. Consequently, although a deterioration in performance is predicted, the dominant function at the origin of such deterioration may very well be one of immediate memory span rather than prosodic segmentation.

5.3. Generalisations

We argue that the neuronal mechanism proposed here – even though tested on digit strings – may be generalised to continuous speech free of linguistic constraints. When listening to everyday speech, intelligibility remains high as long as the delta oscillator is in sync with the input chunk rhythm, in analogy with the now reasonably established observation that, for intelligibility to remain high, the theta oscillator must be in sync with the syllabic rhythm (e.g. Doelling et al., 2014; Ghitza, 2012, 2014). When synchronisation is maintained, a delta cycle is aligned with a speech segment corresponding to a chunk.

Four remarks are relevant here. First, although syllable rate indeed display a peak in the theta range, it varies considerably inside the theta frequency range.¹³ Since the theta oscillator must stay synchronised with the input rhythm, such syllable rate distribution supports the argument for a flexible theta mechanism. Whether rhythm variations at the phrasal level are sufficiently

regular for a flexible delta mechanism to drive a reliable segmentation process is yet to be rigorously quantified.¹⁴ Second, as already reiterated, we hypothesise that in order to secure a reliable performance the acoustic-driven delta oscillator must stay synchronised with the chunk rhythm. If this hypothesis holds a prediction about comprehension can be made, namely that reduced accentuation should lead to reduced comprehension (due to difficulties to stay in sync with the chunk rhythm). Third, a prerequisite for robust tracking is that the cochlear output contains robust information on chunk rhythm, and that there exists a neuronal mechanism – with a flexible oscillator at the core – that can (easily) lock to the chunk rhythm. (One possible mechanism may be a neuronal PLL circuit, for example, Ahissar et al., 1997; Zacksenhouse & Ahissar, 2006.) And fourth, TEMPO suggests that the theta oscillator plays a crucial role in extracting syllable objects (Section 2.1), and that the sequence of syllable objects is integrated to form phrases, delta cycle long (Section 2.2). The need for such coordination may suggest that the theta oscillator is nested within the acoustic-driven delta. Addressing the nature of the nesting – including the consequences of the resulting delta/theta interaction – is beyond the scope of this study.

Recalling the assertions by Miller (1962) – that the duration of the decision unit is about 1-sec long – and by Pickett and Pollack (1963) – that in read passages and in ordinary conversation a window of at least 1 second is required to reliably decode words, irrespective of the number of words presented – we notice that a 1-sec long window corresponds to a 1-Hz oscillation, at the centre of the delta frequency range. We suggest that the emergence of a 1-sec long window as a “sweet spot” is the result of an underlying segmentation mechanism with delta at the core. Noticing that a phrase of about two or three words is about 1-sec long, we further suggest that the commonly reported superior performance for grouping of about three words is also determined by the same segmentation mechanism.

Finally, adopting the view that the strategy of composing words into phrasal units is the result of an evolutionary trajectory to match a cortical function, we hypothesise that the phrasal structure of language is constrained by delta oscillations. Rules of chunking in speech production may be the product of common cortical mechanisms on both motor and sensory sides, with delta at the core. This hypothesis is in line with the hypothesis put forward by Martin (2012).

6. Summary

This study provides psychophysical evidence for the importance of acoustic prosodic segmentation – in

distinction from contextual parsing – in securing a reliable digit retrieval. Importantly, the data show that in order to maintain high level of performance, the phrasal rhythm of the input should be within the delta frequency band (about 0.5 to 3 Hz), giving rise to the possibility of an underlying segmentation mechanism with acoustic-driven delta oscillations at the core. The data show that performance is high for a variety of chunking patterns as long as the chunking rate is inside the delta frequency band, confirming the possibility that chunking strategies of telephone numbers in different languages are of cultural consequence, rather than the result of the need to match a cortical constraint. The data also show that hidden prosody cues – accentuations arching over a chunk – result in grouping with a benefit equivalent to the benefit gained by explicit temporal grouping (i.e. by inserting gaps). We argue that these findings can be generalised to continuous speech free of linguistic constraints, and that the phrase structure of language is constrained by cortical delta oscillations.

Further neuroimaging experiments will be needed in order to validate our hypothesis. No hypothesis about internal physiological processes can be fully validated using only psychophysical methods, and the data reported here establish a psychophysical context for neuroimaging experiments that should use a comparable task.

Then there is the questioning of the brain substructure origins of the hypothesised functions. There is no reason to believe that there is merely one type of delta band response, so there will be no one-size-fits-all answer. Insofar as one can experimentally restrict or control the stimuli to elicit acoustic segmentation at the level a phrase, that contribution to delta would be expected to be primarily associated with auditory cortical regions. Aspects of the delta band response that might be associated with contextual aspects of processing, including working memory, grammatical predictions, etc. are likely to have substantial non-auditory contributions, including from frontal regions. The question will need to be addressed through careful manipulation of the materials, task, and through subtle source reconstruction from neuroimaging data.

Notes

1. The core is a vocabulary of “Lego” speech segments from which the stimuli presented to the listeners were synthesised, by concatenation.
2. A phrase is meant to be a group of words roughly 1-s long – not necessarily a sentence.
3. When attending to time-compressed speech listeners experience insensitivity to moderate time scale

variations; deterioration in intelligibility for compression factors beyond 3; and a recovery of intelligibility by repackaging (e.g. Ghitza, 2014; Ghitza & Greenberg, 2009), where “repackaging” is a process of dividing the time-compressed waveform into fragments, called packets, and delivering the packets in a prescribed rate.

4. Which acoustic landmarks drive the flexible theta oscillator? Two options have been considered, yet to be vetted: (i) CV boundaries (“acoustic edges”), and (ii) vocalic nuclei (“mid vowels”). Here, the vocalic nuclei are preferred because of robustness considerations: in the presence of background noise the “islands” of reliable acoustics are the mid vowel regions (Ghitza, 2013).
5. The theta-syllable (Ghitza, 2013) is a discrete speech-information unit defined by cortical function. Its acoustic correlate is a theta-cycle long speech segment located in between two successive vocalic nuclei. As such, a theta-syllable is aligned with a VCV cluster.
6. This match can be viewed as a synchronisation between the amount of information in the input stream and the necessary decoding time in the pre-lexical level, determined by the flexible theta oscillator (Ghitza, 2011).
7. 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’s internal prosodic rules – and is considered to have some of the finest quality synthesis of any commercial product.
8. The standard deviation here is the square root of the unbiased estimator of the variance.
9. Because these simulations are not simply standard error calculations, the credible intervals are not restricted to be symmetrical around the mean, as can be seen under close inspection of the data later on.
10. 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.)
11. If we were to be consistent with our notation, a 111...11 (a sequence of ten 1’s) should have been used. Alas, we use the shorthand 11111, instead.
12. Time compression uses a pitch-synchronous, overlap and add (PSOLA) procedure (Moulines and Charpentier, 1990) incorporated into PRAAT (<http://www.fon.hum.uva.nl/praat/>) – a speech analysis and modification toolbox. In the time-compressed signal, the formant patterns and other spectral properties are altered in duration; however, the fundamental frequency (pitch) contour remains the same (this is the motivation for using PSOLA methods).
13. This is so across languages.
14. Preliminary data suggest that this indeed is the case (Liberman, 2016a, 2016b; Ryant & Liberman, 2016).

Acknowledgments

I would like to thank *AT&T Labs* and *Interactions LLC* for allowing me access to the Natural Voices Text-To-Speech system,

and in particular to Mark Beutnagel for instructions and advice; to Yair Ghitza for conducting the hierarchical logistic regression analysis of the data; to Nelson Cowan for bringing Ryan’s work (1969) to my attention; to Mark Liberman for sharing the preliminary data on sound/silence durations at the phrasal level; to Peter Cariani and to David Poeppel for commenting on an earlier version of the manuscript; and to the two anonymous reviewers for their thorough comments.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This study was funded by a research grant from the Air Force Office of Scientific Research.

References

- Ahissar, E., & Ahissar, M. (2005). Processing of the temporal envelope of speech. In R. Konig, P. Heil, E. Bauder, & H. Scheich (Eds.), *The auditory cortex. A synthesis of human and animal research* (pp. 295–313, Chap. 18). London: Lawrence Erlbaum.
- Ahissar, E., Haidarliu, S., & Zacksenhouse, M. (1997). Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 11633–11638.
- Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140.
- Boucher, V. J. (2006). On the function of stress rhythms in speech: Evidence of a link with grouping effects on serial memory. *Language and Speech*, *49*(4), 495–519.
- 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. doi:10.1016/j.neuroimage.2008.09.015
- Buzsaki, G., Logothetis, N., & Singer, W. (2013). Scaling brain size, keeping timing: Evolutionary preservation of brain rhythms. *Neuron*, *80*(3), 751–764.
- Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: A reconciliation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*(6), 1235–1249.
- Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*, 341–370.
- Cutler, A. (1994). The perception of rhythm in language. *Cognition*, *50*, 79–81.
- Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.
- Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. doi:10.1038/nn.4186
- Ding, N., & Simon, J. Z. (2009). Neural representations of complex temporal modulations in the human auditory cortex. *Journal of Neurophysiology*, *102*(5), 2731–2743.

- Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual segmentation. *NeuroImage*, *85*, 761–768. doi:10.1016/j.neuroimage.2013.06.035
- Gelman, A. (2005). Analysis of variance – why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New-York, NY: Cambridge University Press.
- Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 130. doi:10.3389/fpsyg.2011.00130
- Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, *3*, 238. doi:10.3389/fpsyg.2012.00238
- Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*, 138. doi:10.3389/fpsyg.2013.00138
- Ghitza, O. (2014). Behavioral evidence for the role of cortical theta oscillations in determining auditory channel capacity for speech. *Frontiers in Psychology*, *5*, 652. doi:10.3389/fpsyg.2014.00652
- 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
- Gilbert, A. C., Boucher, V. J., & Jemel, B. (2014). Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. *Frontiers in Psychology*, *5*(220), 1–9.
- Gilbert, A. C., Boucher, V. J., & Jemel, B. (2015). The perceptual chunking of speech: A demonstration using ERPs. *Brain Research*, *1603*, 101–113.
- Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi:10.1038/nn.3063
- Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, *4*, e06213. doi:10.7554/eLife.06213
- Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, *94*, 1904–1911.
- Liberman, M. (2016a). Political sound and silence. *Language Log*. Feb 8. Retrieved from <http://languagelog.ldc.upenn.edu/nll/?p=23990>
- Liberman, M. (2016b). Poetic sound and silence. *Language Log*. Feb 12. Retrieved from <http://languagelog.ldc.upenn.edu/nll/?p=24054>
- Luce, P. A., & McLennan, C. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 591–609). Malden, MA: Blackwell.
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*, 71–102.
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions during word recognition in continuous speech. *Cognition*, *10*, 487–509.
- Martin, P. (2012). Neurophysiological research explains prosodic structures constraints. *Revista De Estudos Da Linguagem, Belo Horizonte*, *20*(2), 13–22. doi:10.17851/2237-2083.20.2.13-22
- Maybery, M. T., Permentier, F. B. R., & Jones, D. M. (2002). Grouping of list items reflected in the timing of recall: Implications for models of serial verbal memory. *Journal of Memory and Language*, *47*(3), 360–385.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
- Miller, G. A. (1962). Decision units in the perception of speech. *IRE Transactions on Information Theory*, (Feb.), 81–83.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.
- Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Language Science*, *3*, 320. doi:10.3389/fpsyg.2012.00320
- Pickett, J. M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, *6*, 151–164.
- 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
- Reeves, C., Schauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1638–1654.
- Ryan, J. (1969). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, *21*(2), 137–147.
- Ryant, N., & Liberman, M. (2016, September 8–12). Automatic analysis of phonetic speech style dimensions. *Interspeech*, pp. 77–81, San Francisco, CA.
- 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 D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 125–155). Malden, MA: Blackwell.
- Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *Journal of Neuroscience*, *32*(26), 9089–9102.
- Viterbi, A. J. (1966). *Principles of coherent communication*. New York, NY: McGraw-Hill.
- Wickelgren, W. A. (1964). Size of rehearsal group and short-term memory. *Journal of Experimental Psychology*, *68*, 413–419.
- 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.