AFRL-AFOSR-VA-TR-2016-0304

CASCADING OSCILLATORS IN DECODING SPEECH: REFLECTION OF A CORTICAL COMPUTATION PRINCIPLE

Oded Ghitza
TRUSTEES OF BOSTON UNIVERSITY
1 SILBER WAY
BOSTON, MA 02215-1390

09/06/2016
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/RTB2

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-08-2016	Final Performance Report	01 Sep 2011 - 31 Aug 2016

4. TITLE AND SUBTITLE
CASCADING OSCILLATORS IN DECODING SPEECH: REFLECTION OF A CORTICAL COMPUTATION PRINCIPLE

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-11-1-0122

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
DR. 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 02215-1703

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF/AFRL/AFOSR
875 NORTH RANDOLPH STREET, ROOM 3112
ARLINGTON, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

DISTRIBUTION A: Distribution approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Motivated by the possible role of brain rhythms in cortical function, we postulate a cortical computation principle by which decoding 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. Syllabic segmentation utilizes flexible oscillators operating in the theta range (3–9 Hz) by tracking the input syllabic rhythms, and prosodic segmentation is driven by flexible oscillators in the delta range (0.5–3 Hz), tracking prosodic rhythms. A model (TEMPO) was developed which 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. The key properties that enable such accountability are: (i) the capability of the oscillators to track and stay locked to the input rhythm, and (ii) the cascaded nature of the oscillators within the array.

15. SUBJECT TERMS
speech perception, memory access, decoding time, brain rhythms, cascaded cortical oscillations, phase locking, segmentation, parsing, decoding

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			MS. DIANE BALDWIN
dr					19b. TELEPONE NUMBER (*Include area code*) 617-353-4365

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at lest the year and be Year 2000 compliant, e.g., 30-06-1998; xx-08-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. 1F665702D1257.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. AFOSR-82-1234.

5d. PROJECT NUMBER. Enter al project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORS AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-08-2016	Final Performance Report	01 Sep 2011 - 31 Aug 2016

4. TITLE AND SUBTITLE
CASCADING OSCILLATORS IN DECODING SPEECH: REFLECTION OF A CORTICAL COMPUTATION PRINCIPLE

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-11-1-0122

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
DR. 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 02215-1703

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF/AFRL/AFOSR
875 NORTH RANDOLPH STREET, ROOM 3112
ARLINGTON, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Motivated by the possible role of brain rhythms in cortical function, we postulate a cortical computation principle by which decoding 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. Syllabic segmentation utilizes flexible oscillators operating in the theta range (3–9 Hz) by tracking the input syllabic rhythms, and prosodic segmentation is driven by flexible oscillators in the delta range (0.5–3 Hz), tracking prosodic rhythms. A model (TEMPO) was developed which 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. The key properties that enable such accountability are: (i) the capability of the oscillators to track and stay locked to the input rhythm, and (ii) the cascaded nature of the oscillators within the array.

15. SUBJECT TERMS
speech perception, memory access, decoding time, brain rhythms, cascaded cortical oscillations, phase locking, segmentation, parsing, decoding

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			MS. DIANE BALDWIN
dr					19b. TELEPONE NUMBER (*Include area code*) 617-353-4365

EXECUTIVE SUMMARY

This research program aims at understanding how brain circuits enable one of the complex faculties that characterize humans, the communication system. Focusing on speech, the neural and computational principles governing recognition have yet to be formulated, and despite decades of focused research, the way by which humans and machines understand speech remains poorly understood. Motivated by the possible role of brain rhythms in cortical function, we postulate a cortical computation principle by which decoding 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 speech requires that the oscillators be ``flexible'', in contrast to autonomous, ``rigid'' oscillators. Syllabic segmentation utilizes flexible oscillators operating in the theta range (3–9 Hz) by tracking the input syllabic rhythms, and prosodic segmentation is driven by flexible oscillators in the delta range (0.5–3 Hz), tracking prosodic rhythms. Intelligibility is impaired when the ability of the oscillators to synchronize to the input rhythms is impaired. A model (TEMPO) was developed which 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. Ghitza & Greenberg 2009; Ghitza 2012; Doelling et al. 2014; Ghitza, 2014; Ghitza, 2016). The key properties that enable such accountability are: (i) the capability of the oscillators to track and stay locked to the input rhythm, and (ii) the cascaded nature of the oscillators within the array.

Accomplishments:

1. Established the role of theta-driven syllabic parsing in decoding speech by measuring intelligibility of speech with a manipulated modulation spectrum (Ghitza, 2012).
2. Defined the theta-syllable, a unit of speech information defined by cortical function – an alternative to the loosely defined syllable (Ghitza, 2013).
3. Provided behavioral evidence for the role of cortical theta oscillations in determining the capacity of the auditory channel (Ghitza, 2014).
4. Generalized the role of cortical oscillations to music: decoding time for the identification of musical key is determined by brain rhythms (Farbood et al., 2014).
5. Measured whether the excitability spread within a theta cycle affects speech perception. (Pefkou, ongoing PhD thesis).
6. 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, ongoing PhD thesis).
7. Provided neuroimaging (MEG) validation for TEMPO's predictions on the role of acoustic landmarks in driving theta oscillations to facilitate perceptual parsing (Doelling et al., 2014).
8. Developed computational models for components of TEMPO, in MATLAB (Fuglsang, 2015).
9. Provided behavioral evidence for the role of acoustic-driven delta rhythms in setting prosodic markers (Ghitza, 2016).
10. Organized (with David Poeppel) a one-day workshop "Brain Rhythms and Cortical Computation" (BryCoCo). A yearly event.

Publications:

1. Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 3:238. doi:10.3389/fpsyg.2012.00238
2. Ghitza O., Giraud A-L and 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

3. Ghitza O. (2013). "The theta-syllable: a unit of speech information defined by cortical function." Front. Psychol. 4:138. doi: 10.3389/fpsyg.2013.00138
4. Doelling, K. B., Arnal, L. H., Ghitza, O. and Poeppel, D. (2014). "Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing." NeuroImage, 85:761–768. doi: 10.1016/j.neuroimage.2013.06.035
5. Ghitza O (2014) Behavioral evidence for the role of cortical theta oscillations in determining auditory channel capacity for speech. Front. Psychol. 5:652. doi:10.3389/fpsyg.2014.00652
6. 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
7. Fuglsang SA (2015) Towards predicting the intelligibility of time-compressed speech with silence gaps. Centre Applied Hearing Research, Technical University of Denmark. Master thesis. (Supervised by Ghitza O & Dau T).
8. Ghitza O (2016) Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*. A special issue on "Brain oscillations in language comprehension and production". (Accepted.)

Personnel:

1. Ghitza, Oded (BU). PI.
2. Poeppel, David (NYU). Collaboration to accomplishments 4 and 7.
3. Giraud, Anne-Lise (Université de Genève). Collaboration to accomplishment 5.
4. Wingfield, Arthur (Brandeis University). Collaboration to accomplishment 6.
5. Farbood, Morwaread (NYU). Collaboration to accomplishment 4.
6. Dau, Torsten (Technical University Denmark). Collaboration to accomplishment 8.
7. Doelling, Keith (NYU). Graduate student to accomplishment 7.
8. Fuglsang, Søren (Technical University Denmark). Graduate student to accomplishment 8.
9. Pefkou, Maria (Université de Genève). Graduate student to accomplishment 5.
10. Penn, Lana (Brandeis University). Graduate student to accomplishment 6.
11. Rowland, Jess (NYU). Graduate student to accomplishment 4.

Impact:

1. The computation principle embodied in TEMPO advances our understanding of the human cortex, and is directly relevant for improving man-machine interaction.
2. 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.
3. 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*.

1. Background, Motivation

1.1 Speech perception

For nearly a century, the perception of speech has been studied primarily from the perspective of the acoustic signal. Its spectral properties were thought to provide the principle cues for decoding phonemes, and from these building blocks words and phrases would be derived. Early studies (e.g., Fletcher and his Bell Lab colleagues – e.g., Fletcher & Steinberg, 1929, French & Steinberg, 1947), using nonsense syllables, seemed to support the spectral-phonemic model. And indeed, such material, presented in isolation and devoid of lexical, syntactic, and semantic context, does appear under *certain* circumstances to be processed phonemic-segment by phonemic-segment. This perspective seemed so intuitively pleasing and compelling that entire industries (e.g., automatic speech recognition and speech synthesis) were built upon its foundation (see review in Rabiner & Juang 2008; Hinton et al. 2012).

However, a range of phenomena call into question how well the spectral-phonemic model can explain how spoken language is processed in the real world. Some examples to consider include the following:

1) How can we explain the remarkable capability of human listeners to compensate for the large variability in speech acoustics (sometimes summarized as the problem of phonemic invariance and the problem of perceptual constancy)? A specific example is the ability to understand spoken material when time-compressed (e.g., when the signal is compressed by a factor of two intelligibility is almost unaffected). What does this tell us about the relationship between the acoustics and the speech decoding process?
2) What accounts for the specific spectral and temporal properties of spoken language? In particular, why is speech generally spoken at rates between 3 and 7 syllables per second? Does this follow purely from constraints imposed by the motor system, or are there neurophysiological foundations that underlie both perception and production at these rates?
3) Why is speech so robust in the presence of such environmental "distortions" as background noise and reverberation? How do listeners focus on linguistically important components of the signal in the face of an almost infinite variety of degradations?
4) Why is speech so much more comprehensible when spoken in phrases and sentences compared to single words alone? What is there about linguistic context that confers such an advantage in comprehension?

These observations imply that the *underlying representations* are unlikely to be *exclusively* in the form of spectral patterns associated with phonemic (or even lexical) units (Greenberg, 2007). Some other approach is required to understand the resilience and stability of speech perception in real world situations.

In our research program we pursue the hypothesis that a critical component of successfully decoding speech is the process of **perceptual parsing in the time domain.** Parsing is the process by which an input signal is temporally partitioned into temporal units that are (ultimately) linked to a variety of linguistic levels of abstraction, ranging, potentially, from features to phonetic segments to syllables to words and ultimately prosodic phrases. We hypothesize that the parsing process includes the phrasal level (0.5–2 s) and works down to shorter intervals associated with words and syllables (analogous to image-processing algorithms that initially focus on a scene's broad outlines before filling in visual detail, or similarly, 'reverse hierarchies' in cognitive neuroscience models of vision; see Hochstein & Ahissar 2002). **Only after the signal has been parsed can effective decoding proceed.** If the signal is incorrectly parsed, it is more difficult to form a match with internal linguistic patterns associated with segments, words and phrases.

In the past five years we were concerned with the first two questions. Such questions have been addressed in hundreds of studies, usually focused on *specific* aspects of speech. Despite this intensive effort, a coherent, systematic framework for understanding how speech is decoded is still lacking. We believe that

one reason bedeviling such research is its near-exclusive focus on the acoustic aspects of the speech signal. Indeed, such focus has led to a reasonable understanding of the role and function of the auditory periphery (i.e., neural mechanisms responsible for generating a sensory representation), e.g. Young (2008), Mesgarani et al. (2008). In contrast, the research has failed to significantly advance our understanding of the auditory cortical mechanisms involved in decoding speech (at the phonetic, lexical and phrasal levels). This imbalance is reflected in the degree to which models of speech perception can account for observed behavioral data and, in turn, in the way state-of-the-art automatic speech recognition (ASR) systems operate. On the one hand, we have reasonably elaborate models of the auditory periphery (up through the primary auditory cortex) that allow us to design front-ends capable of producing representations that exhibit perceptually important speech information (e.g., Chi et al., 2005; Ghitza et al., 2007; Messing et al., 2009). On the other hand, recognition back-ends are based, almost exclusively, on statistical pattern recognition techniques due to the dearth of data and insight pertaining to cortical processing. Although ASR systems can perform remarkably well under certain conditions, they are able to do so mainly for tasks of limited perplexity.

We adopt the stance that, in order to understand speech decoding by humans and implement it in machines, it is mandatory to incorporate current insights regarding the relevant brain mechanisms (e.g. Hickok & Poeppel 2007; Poeppel et al. 2008). We hypothesize that decoding time is governed by a cascade of neuronal oscillators, locked to the speech rhythm, which guide template-matching operations at a hierarchy of temporal scales. We argue that cascaded cortical oscillations in the theta and gamma frequency bands are crucial for speech intelligibility and that intelligibility is high if these oscillations remain phase-locked to the auditory input rhythm. These computational principles are at the core of a model, TEMPO (Ghitza, 2011), shown to be capable of emulating recent psychophysical data on the intelligibility of speech sentences as a function of syllabic rate (Ghitza and Greenberg, 2009). The data show that intelligibility of speech time-compressed by a factor of 3 (i.e., at a high syllabic rate) is poor (about 50% word error rate), but is substantially restored when silent gaps are inserted in-between successive 40-ms long compressed-signal intervals – a challenging finding, difficult to explain using current models of speech perception, but emerging naturally from the TEMPO architecture. We use TEMPO as a modeling infrastructure, to further investigate the role of cortical oscillatory activity at the psychophysical, neurophysiological and computational levels.

1.2 Cortical oscillations

Neuronal oscillations are believed to play a role in various perceptual and cognitive tasks, including attention (Lakatos et al., 2008), navigation (Buzsaki, 2005), memory (Gruber et al., 2008), motor planning (Donoghue, 1998) and, in the context of the present work, spoken-language comprehension (Bastiaansen & Hagoort 2006; Haarman et al., 2002; Schroeder et al. 2008, de Diego et al. 2011; Peelle et al. 2012). Spatial patterns of neural activation associated with speech processing have been visualized in different regions of the auditory cortex by a variety of brain-imaging methods (PET, fMRI; e.g. Pulvermueller, 1999; Giraud et al., 2004; for review see Hickok & Poeppel 2007, Price 2012). The specific timing of activation across the auditory cortex has been observed with electromagnetic recordings (MEG, EEG, ECoG; e.g. Canolty, 2007; Giraud et al., 2007; Luo & Poeppel, 2007). The modulation of oscillatory activity is typically seen in distinct frequency bands. As will be elaborated, in the context of spoken language comprehension, frequencies of particular relevance are the delta (< 3 Hz), theta (3–8 Hz), beta (15–25 Hz) and gamma (> 30 Hz).

The specific computational functions of neuronal oscillations are uncertain. One possible function is to coordinate the activity of distinct cortical regions and integrate activity across multiple spatial and temporal scales; an oscillatory hierarchy may serve as an organizing instrument for such function (Fries, 2005; von Stein & Sarnthein, 2000; Schroeder et al., 2008). An oscillatory hierarchy may also serve as a central pacemaker, similar to the synchronization facilitator proposed by Singer and others for cortical processing (cf. review by Singer, 1999; Buzsáki, 2006). Such a hierarchy may control neuronal excitability in neuronal ensembles (Lakatos et al., 2005; Kopell & LeMasson, 1994; Hopfield, 2004;

Palva et al., 2005). In particular, these possible functions may play an important role in decoding spoken language.

As previously noted (e.g. Poeppel 2003; Ghitza 2011), there is a remarkable correspondence between average durations of speech units and the frequency ranges of cortical oscillations. Phonetic features (duration of 20–50 ms) are associated with gamma (> 30 Hz) and beta (15–30 Hz) oscillations, syllables and words (mean duration of ~200 ms) with theta (3–8 Hz) oscillations, and sequences of syllables and words embedded within a prosodic phrase (300-1500 ms) with delta oscillations (< 3Hz). In line with this correspondence between cortical oscillations and critical units for the representation of speech, Poeppel (2003) proposed a multi-resolution model where speech is processed concurrently on at least two different time scales (a slow and fast rate), and then information is extracted and combined for lexical access.

A systematic correlation between the acoustics of spoken language and EEG and MEG responses has been demonstrated by showing that temporal cortical responses contain enough information to discriminate single words (Suppes et al., 1997); artificial simple sentences (Suppes et al., 1998), naturalistic sentences (Luo & Poeppel, 2007); and audiovisual speech (Luo et al., 2010), or to correlate with intelligibility (Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et al. 2012). These findings can be interpreted in two distinct ways: (1) cortical oscillations may be part of a *representational* mechanism, with particular oscillations corresponding to modulation frequencies commensurate with intelligible speech; a decoding mechanism may be a phase pattern read-out of theta-band responses, extracted from a sliding 200-ms temporal window – a period of one theta oscillation (Luo & Poeppel, 2007); or (2) the observed cortical rhythms are principally a reflection of an underlying *computational* mechanism, such that the brain sets time intervals for analysis of individual speech components by intrinsic oscillations pre-tuned to an expected speech rate and re-tuned during continuous speech processing by locking to the temporal envelope (Ahissar et al., 2001; Poeppel 2003; Ghitza 2011; Giraud & Poeppel 2012, Ghitza 2012; Doelling et al. 2014; Ghitza 2014). Such a computational principle is in line with the putative role of a hierarchical oscillatory array – controlling neuronal excitability and thus stimulus-related responses in neuronal ensembles (Kopell & LeMasson, 1994; Lakatos et al., 2005; Schroeder et al, 2008; Giraud & Poeppel 2012).

1.3 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. (See: http://hltcoe.jhu.edu/research/challenge-problems/.) In comparing the computation principles utilized by current 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. 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 (embodied in TEMPO), 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 window a 'cortical-time unit', cortical recognition operates in uniform cortical-time units: a theta-syllable object (a VCV) is computed once every theta-cycle (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.

1.4 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 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, 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 five 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 spoken language recognition 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 parsing in recognition which requires verification.

2. Parsing with nested oscillations

Before proceeding further, we define a particular partitioning of the auditory system we shall adhere to:

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

Corollary: The first layer of the *cortical receiver* is the lexical-access circuitry (i.e., words as output).

Such a partitioning of the auditory system stems from the postulate that, when engaging in a spoken dialog, the smallest linguistic meaningful units are words (e.g., Cutler, 1994; Cutler, 2012). According to this partition, syllabic parsing – in charge of distinguishing among syllables – takes place in the auditory channel with processing in time scales in the theta range. Prosodic parsing, which pertains to sequences of words that often contain large amount of information associated with prosodic events, occur in the cortical receiver with processing in time scales in the delta range.

2.1 TEMPO: a model of the auditory channel – pre-lexical (Ghitza, 2011)

In TEMPO (**Fig. 1**), the sensory stream (generated by a model of the auditory periphery, e.g., Chi et al., 2005; Ghitza et al., 2007; Messing et al., 2009) is processed by two concurrent paths (depicted in the figure in blue and orange). The upper (blue) path extracts syllabic-parsing information, which controls the decoding process performed in the lower (orange) path by linking chunks of sensory input with stored memory patterns. Parsing is expressed in the form of an *internal clock-like mechanism* realized as an array of cascaded oscillators, whose frequencies and relative phases determine the processing time frames required for the decoding process.

2.1.1 Parsing

Psychophysical work on the role of temporal modulation (e.g. Dau et al., 1997) and speech modulation in particular (e.g. Houtgast & Steeneken, 1985), demonstrates the relative importance of modulations in the range of 3–9 Hz to intelligibility. This range of modulations turns out to reflect the range of syllable rates in naturally spoken speech, on the one hand (e.g. Pellegrino et al. 2011), and is similar to the frequency range of cortical theta oscillations, on the other. This observation, and the robust presence of these energy

fluctuations in speech acoustics, invites the hypothesis that the *theta* oscillator is the 'master' in the cascaded array, providing syllabic parsing. In accord with neurophysiological data, the core frequency of the theta oscillator in TEMPO is restricted to a frequency range of 3 to 9 Hz. Importantly, in order to be able to track the quasi-regular rhythmicity of speech the oscillators are assumed to be ``flexible'', in contrast to autonomous, ``rigid'' oscillators: we assume that oscillation frequencies are adjusted to optimally match the input syllabic structure by a neuronal version of a phase-lock loop (PLL) system (Viterbi, 1966; Ahissar et al., 1997; Zacksenhouse et al., 2006), where the VCO (voltage controlled oscillator) component is the theta oscillator. The theta oscillator sets the core frequency of the *gamma* oscillator, to be a multiple of the theta frequency. The role of gamma is to determine the time-instances at which the sensory information is sampled within the theta cycle (see Appendix in Ghitza, 2011).

2.1.2 Decoding

In the lower path, the acoustic stream is processed by a template-matching component–a *time-frequency match* component (TFM)– that maps phonetic primitives to memory neurons, termed *VCV neurons*, by computing coincidence in firing across auditory (i.e., tonotopic) frequency channels. At this level, time-frequency patterns are matched over syllable-long time intervals (about 250 ms). These are often formant transitions associated with such phonetic features as place of articulation, which is important for distinguishing consonants (and hence words). This operation is performed within a theta cycle, and mapping onto a VCV neuron occurs at the end of each theta cycle. The role of gamma is different: it determines the time-instances at which the sensory information is sampled within the theta cycle. One possible realization of the TFM component is a version of a model suggested by Shamir et al. (2009). It is a model for the representation of time-varying stimuli (e.g. speech syllables) by a network exhibiting oscillations on a faster time scale (e.g. gamma). An important property of the model is the *insensitivity to time-scale variations* (see Appendix in Ghitza, 2011).

2.1.3 Psychophysical plausibility

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. The data, collected during the past 5 years, is presented below, in Sec. 3. 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)[1]. 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). The cascaded oscillatory array possesses three properties that prove to be crucial for successful account of psychophysical data. Two are inspired by solid findings of the characteristics of cortical oscillations: (a) each oscillator has a finite range of locking oscillation frequencies; (b) oscillators in the array are related (by nesting, e.g. Schroeder & Lakatos 2008; Malerba & Kopell 2013). The third property emerges from a hypothesis, central to TEMPO: (c) the oscillators are capable of remaining locked to the input syllabic rhythm as its slowly changes with time (Ahissar et al. 2001; Nourski et al. 2009).

2.1.4 Neurophysiological plausibility: phase locking and nesting (Giraud & Poeppel 2012)

Fig. 2 (from Giraud & Poeppel 2012) illustrates a neurophysiological model, parallel to TEMPO in most respects, of some of the hypothesized early processing steps. A series of experiments suggests that intrinsic neuronal oscillations in cortex at 'privileged' frequencies (delta 1–3 Hz, theta 3–9 Hz, low

[1] This match can be viewed as a synchronization 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).

gamma 30-50 Hz) may provide some of the relevant mechanisms. To achieve parsing of a naturalistic input signal (e.g. speech signal on top) into elementary pieces, one 'mesoscopic-level' mechanism is suggested to be the sliding and resetting of temporal windows, implemented as phase locking of low-frequency (delta, theta) activity to the envelope of speech and phase resetting of the intrinsic oscillations on privileged time scales. The successful resetting of neuronal activity, triggered in part by stimulus-driven spikes, provides time constants (or temporal integration windows) for parsing and decoding speech signals. Recent studies link the infrastructure provided by neural oscillations (which reflect neuronal excitability cycles) to principled perceptual challenges in speech recognition.

The potential role of neuronal oscillations for speech processing is a rich and highly controversial area of current research (Peelle & Davis 2012). It is imperative to improve our understanding, as the ideas have already penetrated the clinical literature, in particular in recent proposals on the origins of dyslexia (Goswami 2011; Power et al. 2012; Lehongre et al. 2011). Two major issues of broad interest for basic and clinical questions concern (i) the functional role of low-frequency delta and theta oscillations and their potential interrelationship, as well as their causal contribution to intelligibility; and (ii) what mechanisms mediate signal tracking and nesting. The promise of our research program is the potential to develop mechanistic linking hypotheses between speech processing and theories of neural coding.

3. Accomplishment during past 5 years

3.0 Prelude. Prior research (AFOSR supported)

Accomplished prior to the time period covered by this grant. It is briefly reviewed here because of its relevance to TEMPO and to the accomplishments outlined below.

Intelligibility of time-compressed speech with insertion of silence gaps (Ghitza & Greenberg, 2009). In the context of various temporal manipulations, time-compression studies provide valuable insight into how the brain decodes speech. Intelligibility suffers little, as long as the compression/expansion ratio is less than three. This observation is interesting because it demonstrates that the relation between speech acoustics and internal representations is complex and non-linear. Radical time compression reveals how complex speech decoding really is. In the study, the intelligibility of naturally spoken, semantically unpredictable sentences (i.e., without context) time-compressed by a factor of 3, with insertions of silent gaps in-between successive intervals of the compressed speech was measured. **Fig. 3** depicts the critical result. Without insertions, intelligibility was poor (about 50% word error rate); but it was restored considerably by the insertion of gaps, as long as the gaps were between 20 and 120 ms. Since the duration of the acoustic sample (or glimpse) was held constant (40 ms) the sole varying parameter was the length of the inserted gap, hence any change in intelligibility could be attributed to the length of the inserted gap *per se*, rather than to the amount of information contained in the acoustic interval. No (purely) auditory or articulatory model can explain this behavior. The insertion of gaps was interpreted as the act of providing extra decoding time (a *cortical* factor) via "re-packaging" the information stream. Furthermore, it was hypothesized that decoding time is governed by low-frequency brain oscillations. Maximal perceptual restoration occurred when the gaps were 80-ms long (equivalent to a packaging rate of 8.3 packets/sec). The range of values at the bottom of the U-shape performance curve in Figure 3 is roughly equivalent to the theta range.

3.1 Psychophysical results

The role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum Ghitza, 2012). The hypothesized role of theta was then examined by measuring the intelligibility of speech with a manipulated modulation spectrum. Each critical-band signal was manipulated by controlling the degree of temporal envelope flatness. The material comprised 100 7-digit strings spoken fluently by a male speaker. Intelligibility was measured in terms of digit error rate and string error rate. Intelligibility of speech with critical-band envelopes that are flat is poor, as shown in **Fig. 4**. Inserting extra information, restricted to the input syllabic rhythm, markedly improves intelligibility. It

is concluded that flattening the critical-band envelopes prevents the theta oscillator from tracking the input rhythm, hence the disruption of the hierarchical window structure that controls the decoding process, hence the intelligibility decline. Reinstating the input-rhythm information revives the tracking capability, hence restoring the synchronization between the window structure and the input, resulting in the extraction of additional information from the flat modulation spectrum.

The theta-syllable: a unit of speech information defined by cortical function (Ghitza, 2013). This study proposed an alternative to the conventional, ambiguously defined syllable. The **theta-syllable** is defined as a theta-cycle long speech fragment located in between two successive vocalic nuclei. During a successful tracking by the theta oscillator (for uncompressed speech, in quiet, this is the normative case), one theta-cycle is aligned with the acoustics between two successive vocalic nuclei. As such, the theta-syllable is a non-ambiguous acoustic correlate to a VCV (the C stands for consonant cluster). Given the prominence of vocalic nuclei in the presence of environmental noise, theta-syllable is robustly defined. It is also invariant to time scale modifications that result in intelligible speech. When listening to time-compressed speech that is intelligible, the cortical theta is in sync with the stimulus. Because of the cascading nature of the oscillatory array, as long as synchronization holds, the acoustics and the sampling points—determined by gamma oscillations—maintain the same relationship (they "breathe" together), resulting in the same neural code for the underlying VCV.

The role of cortical theta oscillations in determining auditory channel capacity (Ghitza, 2014). Auditory channel capacity of time-compressed speech was measure by introducing "repackaging" – a process of dividing the time-compressed waveform into fragments, called packets, and delivering the packets in a prescribed rate. For all compression factors tested (up to eight), packaging rate at capacity equals 9 packets/sec – aligned with the upper limit of cortical theta, θ_{max} (about 9 Hz) – and the packet duration equals the duration of one uncompressed theta-syllable divided by the compression factor. The alignment of both the packaging rate and the packet duration with properties of cortical theta suggests that the auditory channel capacity is determined by theta. Irrespective of speech speed, the maximum information transfer rate through the auditory channel is the information in one uncompressed theta-syllable long speech fragment per one θ_{max} cycle. Equivalently, the auditory channel capacity is 9 theta-syllables/sec. **Fig. 5** shows the packaging rate $\boldsymbol{\varphi}^*$ ($\boldsymbol{\varphi}^* = \theta_{max}$) and the packet duration δ^o at capacity, for various compression factors ($\boldsymbol{\kappa}$). For every $\boldsymbol{\kappa}$, the spectro-temporal information carried by a packet (i.e., the uncompressed theta-syllable) is the same, albeit time-compressed.

Acoustic-driven delta rhythms as prosodic markers (Ghitza, 2016). This study provides psychophysical evidence for the importance of acoustic prosodic segmentation—in distinction from contextual parsing—in securing a reliable digit retrieval. The experiments used context-free random-digit strings in order to focus exclusively on bottom-up processes. Two experiments were 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. As is shown in **Fig. 6,** 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 (about 0.5 to 3 Hz), but sharply deteriorates for higher rates, 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 generalized to continuous speech free of linguistic constraints, and that the phrase structure of language is constrained by cortical delta oscillations.

Decoding time for the identification of musical key (Farbood et al., 2014). This psychophysical study examines the decoding time at which the brain processes structural (key) information in music and

compares them to timescales implicated in recent work on speech. Combining an experimental paradigm based on Ghitza and Greenberg (2009) for speech with the approach of Farbood et al. (2013) to timing in key-finding, listeners were asked to judge the key of short melodic sequences that were presented at varying tempi, with varying durations of silence inserted in a periodic manner in the audio signal. The distorted audio signals comprised of signal-silence alternations show error rate curves that identify peak performance centered around an event rate of 5–7 Hz (143–200 ms inter-onset interval; 300–420 beats per minute), where event rate is defined as the average rate of pitch change. **Fig. 7** shows error rate plotted by event rate. The results reveal that the error rate minimum centered around 5–7 Hz for all audio segment durations. The data support the hypothesis that the perceptual analysis of music encompasses the processes of parsing the signal into chunks of the appropriate temporal granularity and decoding the signal for recognition. The music-speech comparison points to similarities in how auditory processing builds on the specific temporal structure of the input, and how that structure interacts with the internal temporal dynamics of the neural mechanisms underpinning perception.

Excitability spread within a theta cycle -- does it affect speech perception? (Pefkou et al., in preparation). According to TEMPO, decoding is the result of sampling the acoustics by the gamma, nested in theta (TFM, Section 2.1.2); it is also assumed that the gamma cycles are evenly spread inside a theta cycle. Here we examine this hypothesis and ask if the excitability spread within a theta cycle is uniform. For a time-compressed speech undergone repackaging the packet occupies only part of the packaging cycle; in Ghitza & Greenberg (2009) and Ghitza (2014), the packet was in a fixed position, aligned with the start of the packaging cycle. In **Fig. 8** each row represents a spectrogram of a signal, time-compressed and repackaged. In all rows the compression factor **κ,** the packaging rate **φ**, and the packet duration **δ** are the same (see definition of **φ**, **Δ** and **δ** in the legend of **Fig. 8)**. However, the packet is split into two sub-packets identical in duration. Intelligibility is measured as a function of the location of the 2nd sub-packet. How does the location of 2nd sub-packet affect performance?

Aging and fast speech: decline in performance due to slower theta? (Penn et al., in preparation). The performance of older adults in perceiving speech is affected much more by an increase in speech speed compared to that of young adults (e.g., Gordon-Salant and Fitzgibbins, 1993; Wingfield et al., 1999), and this can be attributed to both sensory deficits and cognitive slowing. This study tests the hypothesis that cognitive slowing plays a major role in this deficit. The present experiments use context-free, random-digit strings in order to eliminate the effect of contextual information. **Fig. 9** shows the performance of Young adults (blue) and Elderly (red) in a digit recognition task. Two stimulus conditions were used: (i) speech undergone uniform time compression (using PSOLA[2]), in thin lines; and (ii) time-compressed speech undergone repackaging, similar to Ghitza & Greenberg, 2009 (reviewed here in the preview to section 3; see updated definition of repackaging in Ghitza, 2014), in thick lines. The acoustics inside the packet is same as the acoustics in condition (i), i.e., same speed; and the packet duration is 20% of the packaging cycle (i.e., 20% duty cycle). We define the **crossover point** (arrows in **Fig. 9**) as the point beyond which repackaging results in intelligibility restoration. For speech uniformly compressed, performance declines with the increase of speed (thin lines), with the crossover point at a rate of 14 syllables/sec for the Young and 11 syllables/sec for the Elderly. Below the crossover point subjects can deal with fast speed, albeit a decline in performance. When speech is repackaged, the crossover point is at 4.5 packets/sec for the Young and 3.7 packets/sec for the Elderly. Above the crossover point performance for the repackaged speech is better than that of the uniformly compressed speech, i.e., intelligibility is partially restored. Interpreting this data through the prism of TEMPO we suggest: (i) intelligibility restoration is because repackaging brings information transfer speed into theta range; and (ii) the fact that

[2] Time compression uses a pitch-synchronous, overlap and add (PSOLA) procedure (Moulines & 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).

the crossover point of the Elderly is lower than that of the Young indicates a shift of the theta frequency range downwards.

3.2 Neurophysiological results

Acoustic landmarks drive delta–theta oscillations to facilitate perceptual parsing (Doelling et al., 2014). This study builds on the recent behavioral study, discussed in Section 3.1, where temporal features of speech are manipulated in order to delineate the role of temporal syllabic cues in speech intelligibility (Ghitza 2012). Removing temporal fluctuations in the envelope that relate to syllabic rate (2–9 Hz) significantly reduced the intelligibility of the degraded stimulus (**Fig. 4**). Surprisingly, when noise clicks were added to the stimulus where the peaks in the original envelope would have been, the error rate dropped by about 50%. The proposed interpretation is that removing these cues disrupts the ability of cortical theta to track the envelope. Whereas reducing the rhythmic structure of the stimulus reduced intelligibility, reinstating that rhythm artificially using short temporal spectro-temporal markers presumably enhanced tracking – and thus intelligibility. We hypothesize that tracking is driven by these relatively sharp temporal fluctuations that relate to the syllabic rate (cf. Howard & Poeppel 2010, 2012). Specifically, we propose that large temporal fluctuations in the stimulus envelope – occurring at a rate within the syllabic-rate range – constitute auditory landmarks that reset the phase of intrinsic cortical oscillation at the theta range. These phase resets generate the envelope tracking behavior that parses the stimulus into syllable-size representations. Using MEG recordings, the 'cerebro-acoustic coherence' (a measure of tracking, Peelle et al. 2012) was calculated and then correlated that metric with intelligibility. Results show that by removing temporal fluctuations that occur at the syllabic rate, envelope-tracking activity is reduced. By artificially reinstating these temporal fluctuations, envelope-tracking activity is regained. **Fig. 10** shows a robust positive correlation between neural tracking behavior and intelligibility from 2.5–4 Hz is observed. This range precisely matches the average syllabic rate of these stimuli (~3 Hz). This suggests that envelope tracking at the dominant (syllabic) rate, while not indicative of intelligibility on the whole, is a prerequisite. Interestingly, our preliminary results also show an unexpected negative correlation in the alpha range (9.5–12 Hz). This may relate to findings suggesting that in degraded stimuli, alpha activity may perform functions pertaining to working memory and cognitive load, and as such may be independent of the stimulus (Obleser et al. 2012). These changes in tracking correlate with intelligibility of the stimulus. Together, the results suggest that the fluctuations in the stimulus, as reflected in the cochlear output, drive oscillatory activity to track and entrain to the stimulus, at its syllabic rate. This process likely facilitates parsing of the stimulus into meaningful chunks appropriate for subsequent decoding, enhancing perception and intelligibility.

3.3 Algorithmic implementation

Computational models for components of TEMPO (Fuglsang, 2015). A model was developed which uses modulation spectrograms to construct an oscillating time-series synchronized with the slowly varying input rhythm. **Fig. 11** illustrates how this model could be used to realize the parsing path of TEMPO. The upper panel shows a waveform of a 10-digit string '105 865 82 63' – a telephone number uttered in American English pronunciation – with a 160 ms gap between every chunk of digits. The middle panel shows the cochlear representation of the waveform. The bottom panel shows the output of the automatic parser. The time-interval between two successive red markers stands for a delta cycle and represents **prosodic** parsing. Similarly, the time-interval between two successive blue markers stands for a theta cycle and represents **syllabic** parsing. (The light red and light blue traces, from which the markers are extracted from, are derived from the modulation spectrum, not shown here.) In another part of the project, not discussed here, different strategies for implementing the decoding path are discussed, and a possible realization of the decoding path is presented. The findings of this study serve as an initial step towards developing a TEMPO-based word recognition system.

3.3 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 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) 112 times and 349 times, respectively. The foundational Ghitza & Greenberg (2009) psychophysics paper on time-compressed speech has been cited 102 times; and two very recent papers in which we jointly test the causal role of theta-entrainment in a "manipulated modulation spectrum" (Ghitza, 2012; Doelling et al. 2014) have already been cited 34 and 58 times, respectively. A new paper that extends our approach to the delta band and incorporates phrase- and sentence-level processing was just published (Ding et al. 2015) but has received considerable press attention. 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., and Ahissar, M. (2005). "Processing of the temporal envelope of speech," in *The Auditory Cortex. A Synthesis of Human and Animal Research*, Chap. 18, eds R. Konig, P. Heil, E. Bundinger and H. Scheich (London: Lawrence Erlbaum), 295–313.

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.

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98:13367–13372.

Bastiaansen M, Hagoort P (2006) Oscillatory neuronal dynamics during language comprehension. *Prog Brain Res* 159:179–196.

Buzsáki G (2005) Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus* 15(7):827–840.

Buzsáki G (2006) *Rhythms of the Brain*. Oxford University Press, New York.

Canolty RT, Soltani M, Dalal SS, Edwards E, Dronkers NF, Nagarajan SS, Kirsch HE, Barbaro NM, Knight RT (2007) Spatiotemporal dynamics of word processing in the human brain. *Frontiers in Neurosciences* 1(1):185–196.

Chi T, Ru P, Shamma SA (2005) Multi-resolution spectro-temporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118: 887−906

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.

Dau T, Kollmeier B, Kohlrausch A (1997) Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.* 102(5):2892−2905.

de Diego-Balaguer R, Fuentemilla Ll, Rodriguez-Fornells A. (2011) Brain dynamics sustaining rapid rule extraction from speech. *Journal of Cognitive Neuroscience* 23(10):3105-20.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2015) Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neurosci*. doi:10.1038/nn.4186

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. doi:10.1016/j.neuroimage.2013.06.035

Donoghue JP, Sanes JN, Hatsopoulos NG, Gaál G (1998) Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements. *J Neurophysiol* 79(1):159–173.

Farbood MM, Marcus G, Poeppel D (2013) Temporal dynamics and the identification of musical key. *Journal of Experimental Psychology: Human Perception and Performance*, 39(4), 911–918. doi:10.1037/a0031087

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.

Fuglsang SA (2015) Towards predicting the intelligibility of time-compressed speech with silence gaps. *Centre Applied Hearing Research, Technical University of Denmark*. Master thesis. (With Dau T).

French NR, Steinberg JC (1947) Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* 19:90-119.

Fries P (2005) A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends. Cogn. Sci.* 9(10):474−480.

Ghitza O, Messing D, Delhorne L, Braida L, Bruckert E, Sondhi MM (2007) Towards predicting consonant confusions of degraded speech. In: Hearing – from sensory processing to perception (Eds.) B Kollmeier, G Klump, V Hohmann, U Langemann, M Mauermann, S Uppenkamp, J Verhey, Springer-Verlag, Berlin Heidelberg, 541–550.

Ghitza O (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol*, 2:130.

Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front. Psychol*. 3:238. doi:10.3389/fpsyg.2012.00238

Ghitza O (2013) The theta-syllable: a unit of speech information defined by cortical function. *Front. Psychol*. 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. *Front. Psychol*. 5:652. doi:10.3389/fpsyg.2014.00652

Ghitza O (2016) Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*. A special issue on "Brain oscillations in language comprehension and production". (Accepted.)

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.

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

Giraud AL, Kell C, Thierfelder C, Sterzer P, Russ MO, Preibisch C, Kleinschmidt A (2004) Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex* 14:247–255.

Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H (2007) Intrinsic cortical rhythms determine cerebral specialization for speech perceptionand production. *Neuron* 56:1–8.

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15(4): 511-7.

Gordon-Salant S, Fitzgibbons PJ (1993) Temporal factors and speech recognition performance in young and elderly listeners. *J. Speech Hearing Res.* 36, 1276–1285.

Goswami U (2011) A temporal sampling framework for developmental dyslexia. *Trends Cogn Sci*. 15(1):3-10.

Greenberg S (2007) What makes speech stick*? Proceed. Intern. Conf. on Phonetics*, ICPhS XVI.

Gruber T, Tsivilis D, Giabbiconi CM, Müller MM (2008) Induced electroencephalogram oscillations during source memory: familiarity is reflected in the gamma band, recollection in the theta band. *J Cogn. Neurosci.* 20:1043–1053.

Haarman HJ, Cameron JA, Ruchkin DS (2002) Neuronal synchronization mediates on-line sentence processing: EEG coherence evidence from filler-gap constructions. *Psychophysiology* 39:820−825.

Hickok G, Poeppel D (2007) *The cortical organization of speech processing.* Nat Rev Neurosci. 8(5):393-402.

Hinton G, Deng l, Yu D, Dahl GE, Mohamed A-r, Jaitly N, Senior A, Vanhoucke v, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82-97. doi :10.1109/MSP.2012.2205597

Hochstein S, Ahissar, M (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36 (5): 791-804.

Hopfield JJ (2004) Encoding for computation: Recognizing brief dynamical patterns by exploiting effects of weak rhythms on action-potential timing. *PNAS*, 101(16): 6255-6260.

Houtgast T, Steeneken HJM (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* 77:1069–1077.

Kopell N, LeMasson G (1994) Rhythmogenesis, amplitude modulation and multiplexing in a cortical architecture'.*Proc. Nat. Acad. Sci. USA* 91:10586−10590.

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

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320:110–113.

Lehongre K, Ramus F, Villiermet N, Schwartz D, Giraud AL (2011) Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. *Neuron*. 72(6): 080-90.

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54:1001–1010.

Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am*. 123(2):899−909.

Messing DP, Delhorne L, Bruckert E, Braida 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

Moulines E, Charpentier F (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.

Nourski KV, Reale RA,Oya H, Kawasaki H,Kovach CK, Chen H, Howard III MA, Brugge JF (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience* 29(49):15564–15574.

Obleser J, Herrmann B, Henry MJ (2012) Neural Oscillations in Speech: Don't be Enslaved by the Envelope. Front Hum Neurosci. 6:250.

Palva JM, Palva S, Kaila K (2005) Phase synchrony among neuronal oscillations in the human cortex. *Journal of Neuroscience* 25(15):3962–3972.

Peelle JE, Gross J, Davis MH (2012) Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. Cereb Cortex.

Peelle JE, Davis MH (2012) Neural Oscillations Carry Speech Rhythm through to Comprehension. *Front Psychol*. 3:320.

Pefkou M (2015) Excitability spread within a theta cycle: does it affect speech perception? *Department of Neuroscience, University of Geneva*. PhD thesis, work in progress. (With Giraud AL)

Pellegrino F, Coupé C, Marsico E (2011) A cross-*language* perspective on speech information rate. *Language*. 87(3): 539-558.

Penn L (2015) Aging and compressed speech: decline in performance due to slower theta? *Memory and Cognition Lab, Brandeis University*. PhD thesis, work in progress. (With Wingfield A)

Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication* 41:245–255.

Poeppel D, Idsardi W, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond B* 363:1071−86.

Power AJ, Mead N, Barnes L, Goswami U (2012) Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. *Front. Psychology* 3:216.

Price CJ (2012) A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage* 62(2), 816-847

Pulvermueller F (1999) Words in the brain's language. *Behavior and Brain Science* 22:253−366.

Rabiner L, Juang BH (2008) Pare E: Speech Recognition. In *Handbook of Speech Processing* (eds. Benesty J, Sondhi MM, Huang Y) 413-428. Berlin: Springer-Verlag.

Räsänen O, Doyle G, Frank MC (2015) Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *INTERSPEECH*. pp. 3204–3207

Schroeder CE, Lakatos P (2008) Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences.* 32(1):9–18.

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(5):1–12.

Singer W (1999) Neuronal Synchrony: A versatile code for the definition of relations? *Neuron* 24(1):49–65.

von Stein A, Sarnthein J (2000) Different frequencies for different scales of cortical integration: from local gamma to long range alpha / theta synchronization. *International Jour. Psychophysiology* 38:301−313.

Suppes P, Lu ZL, Han B (1997) Brain-wave representation of words. *Proc. Natl. Acad. Sci. USA* 94:14965–14969.

Suppes P, Han B, Lu ZL (1998) Brain-wave representation of sentences. *Proc. Natl. Acad. Sci. USA* 95:15861 –15866.

VanRullen R, Zoefel B, Ilhan B (2014). On the cyclic nature of perception in vision versus audition. *Phil. Trans. R. Soc. B*, 369 (20130214).

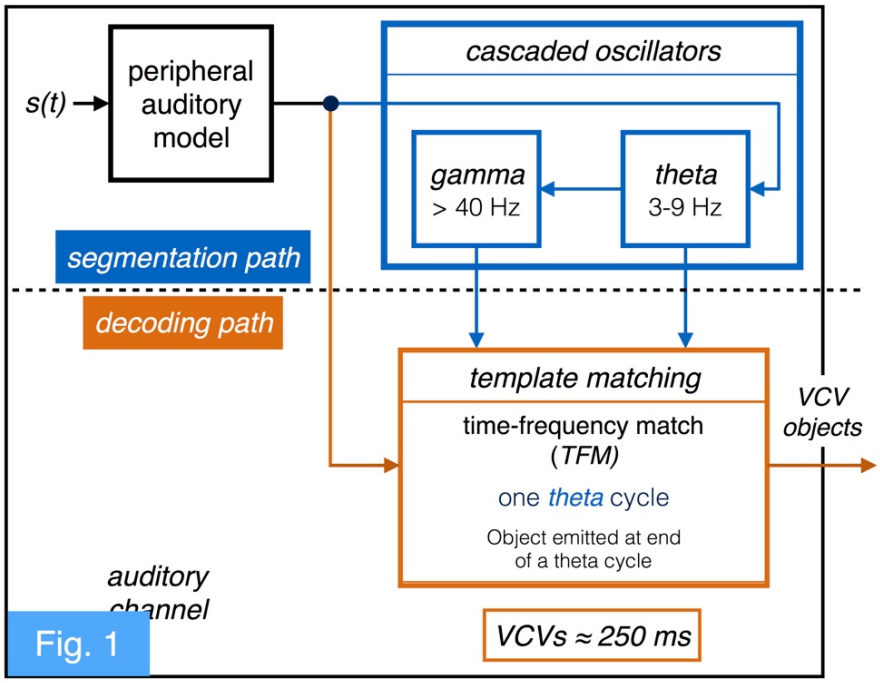
Viterbi AJ (1966) *Principles of coherent communication*. McGraw-Hill, New York.

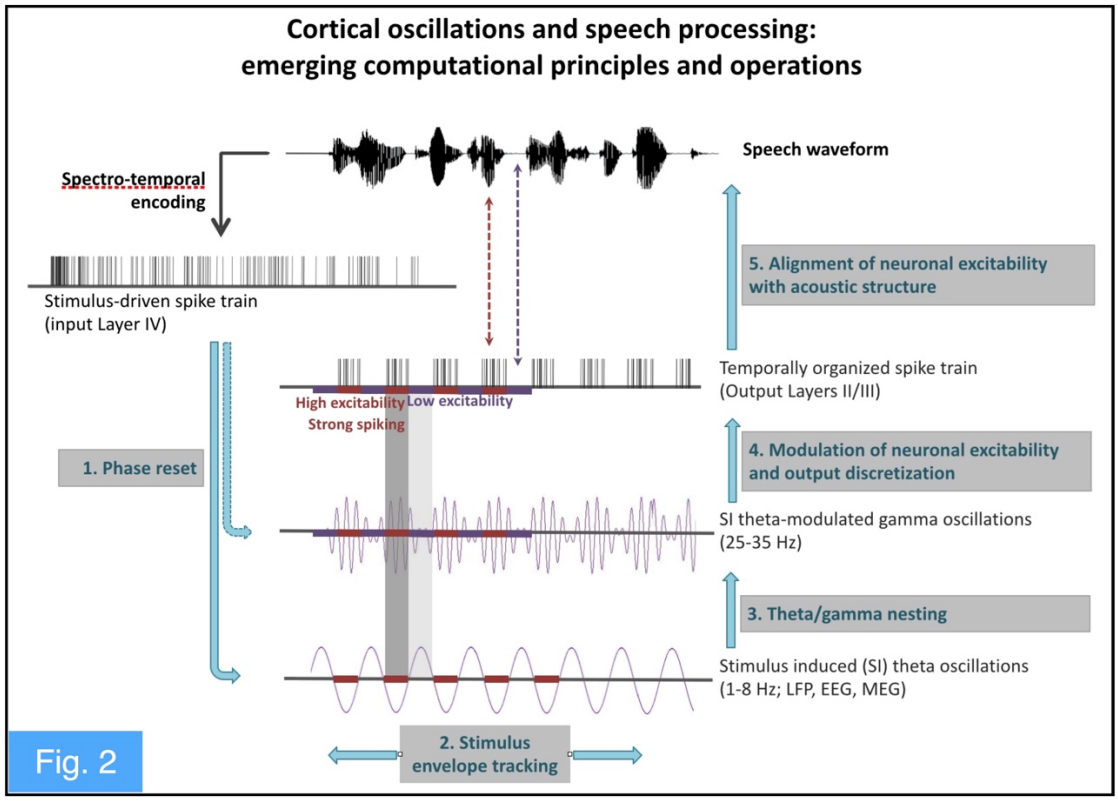
Wingfield A Tun PA, Koh CK, Rosen MJ (1999) Regaining lost time: adult aging and the effect of time restoration on recall of time-compressed speech. *Psychology and Aging* 14, 380−389.

Young ED (2008) Neural representation of spectral and temporal information in speech. *Philos Trans R Soc* 363(1493):923–45.

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.

Figures





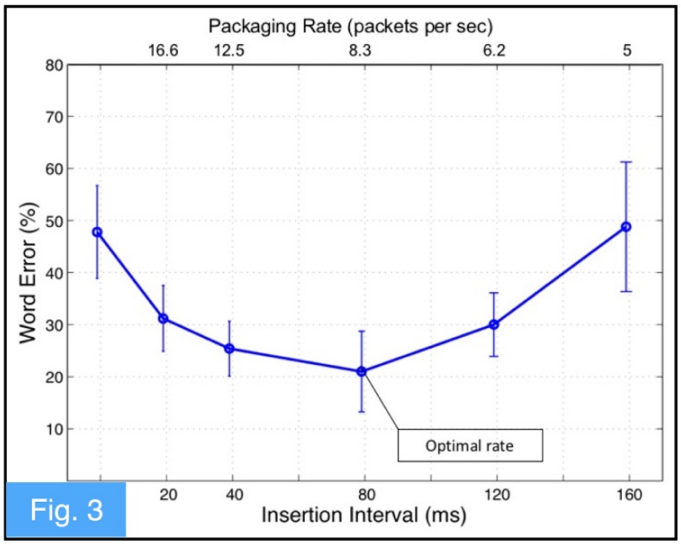


Fig. 3

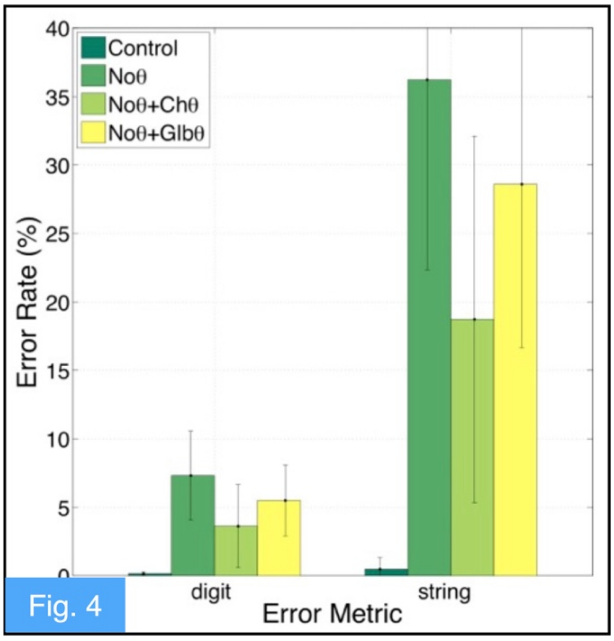


Fig. 4

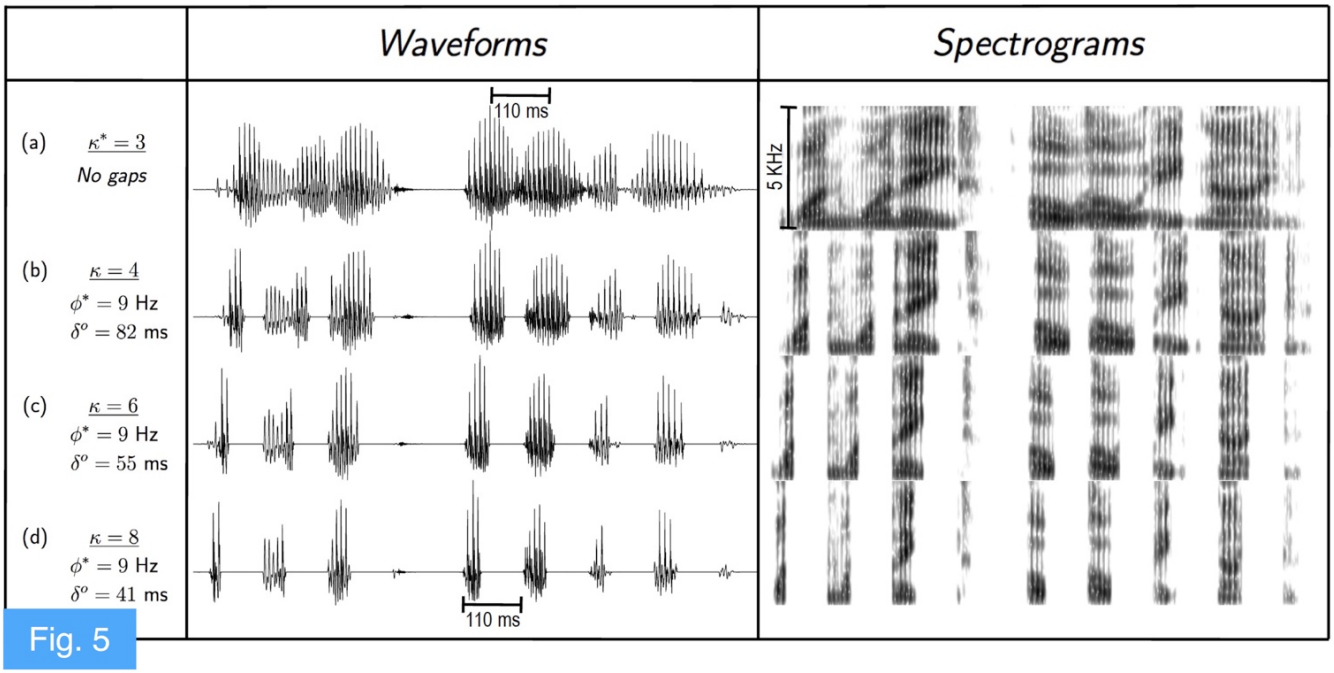


Fig. 5

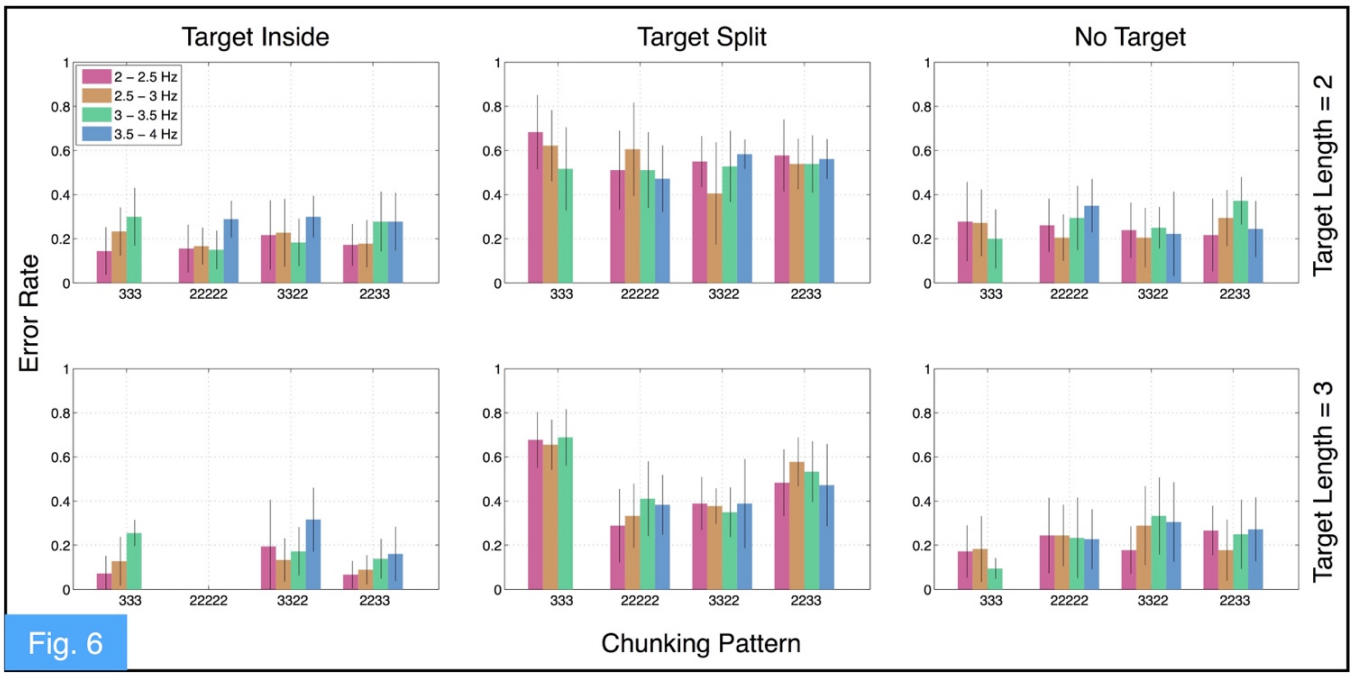


Fig. 6

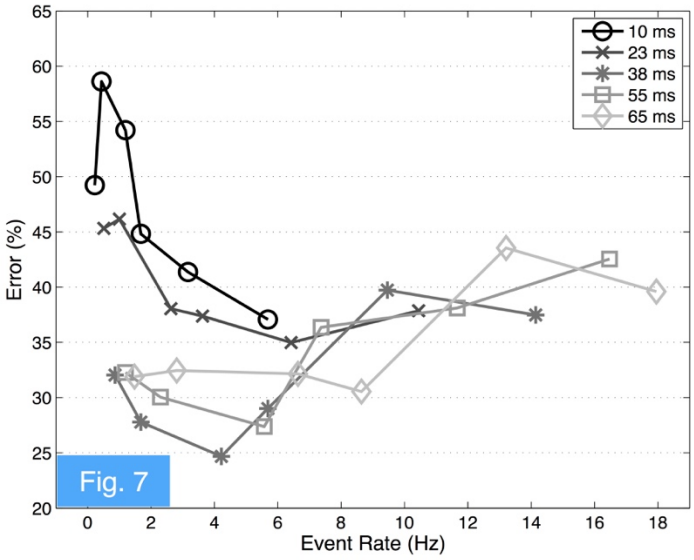


Fig. 7

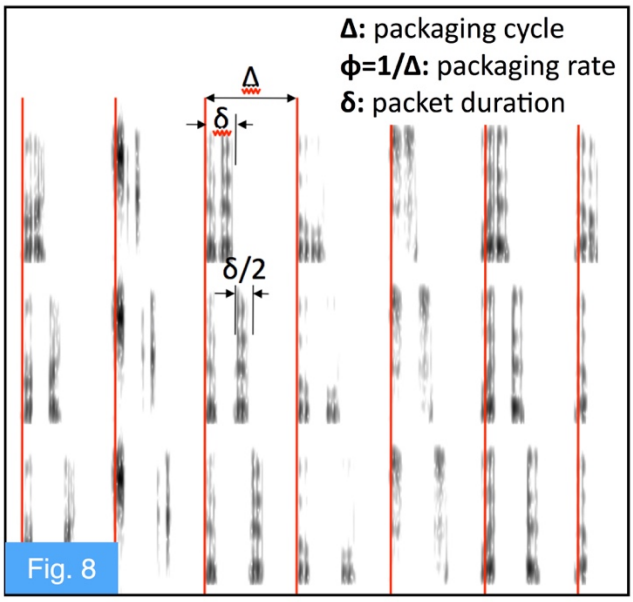


Fig. 8

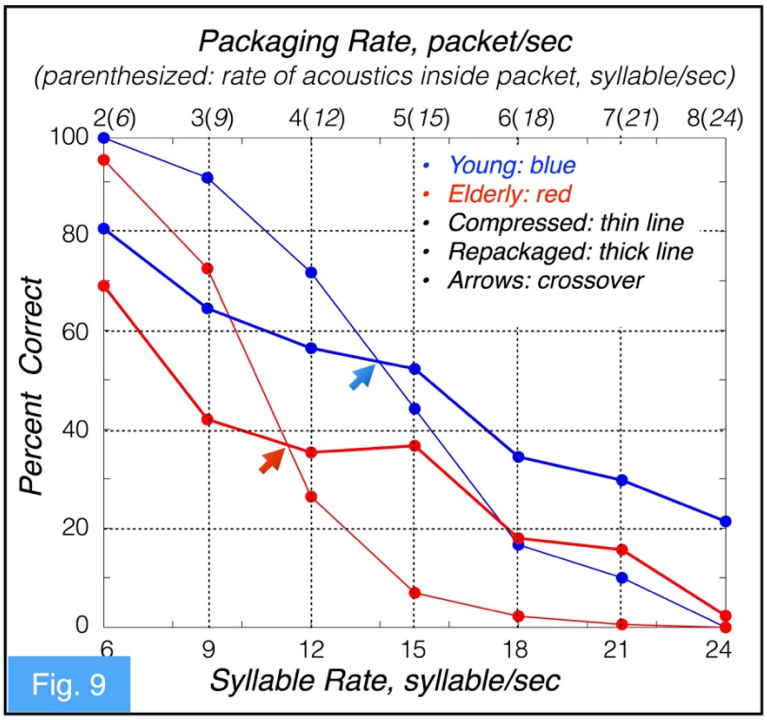


Fig. 9

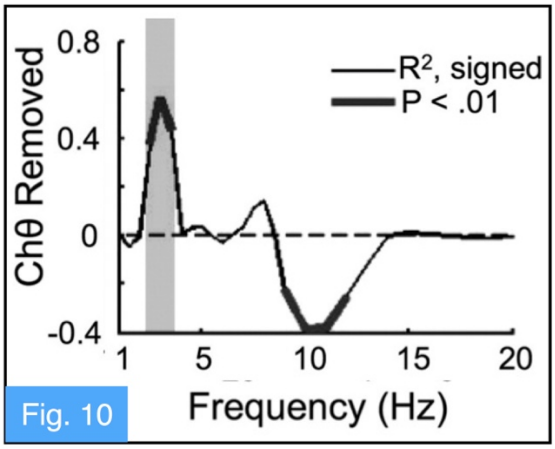


Fig. 10

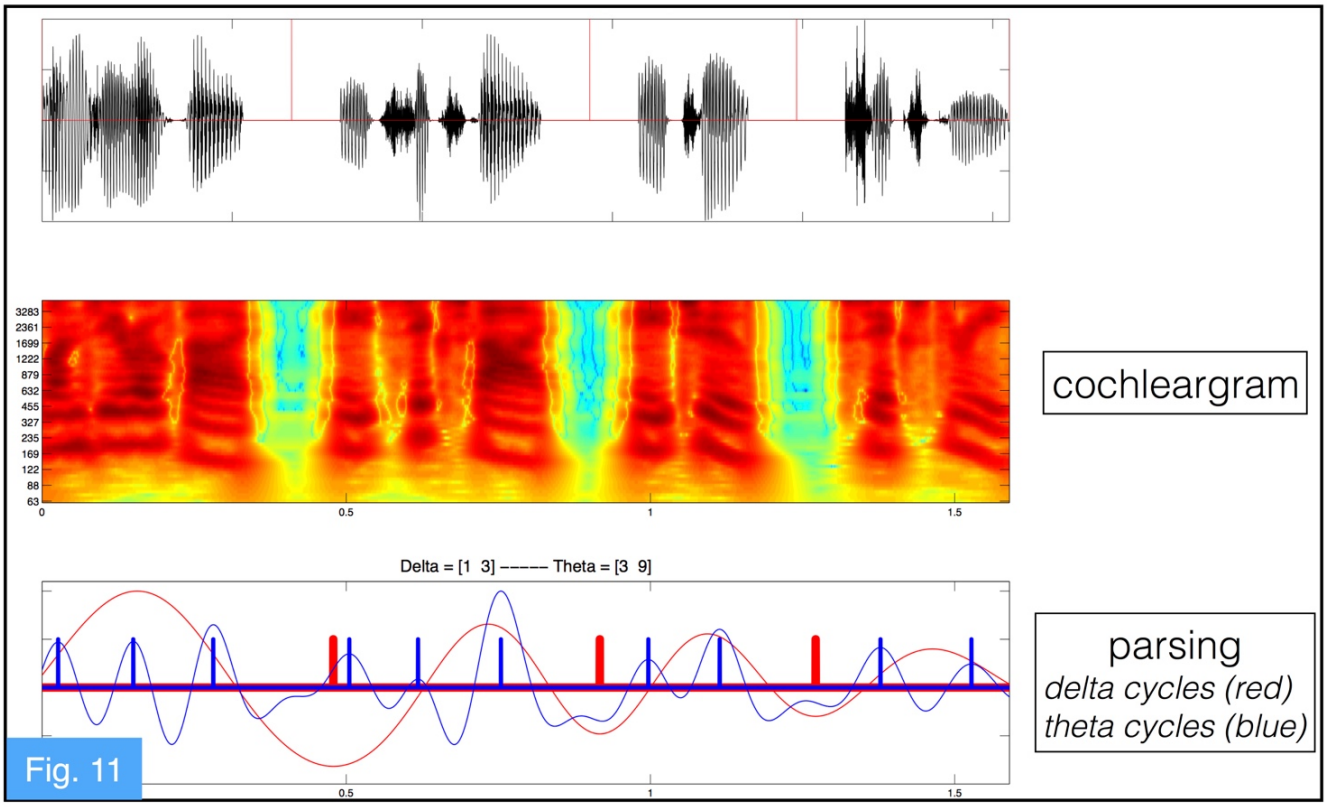


Fig. 11

AFOSR Deliverables Submission Survey

Response ID:6790 Data

1.

Report Type

Final Report

Primary Contact Email

Contact email if there is a problem with the report.

oghitza@bu.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

617-358-1948

Organization / Institution name

Boston University

Grant/Contract Title

The full title of the funded effort.

Cascading oscillators in decoding speech: Reflection of a cortical computation principle

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-11-1-0122

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Oded Ghitza

Program Officer

The AFOSR Program Officer currently assigned to the award

Patrick.Bradshaw

Reporting Period Start Date

09/01/2011

Reporting Period End Date

08/31/2016

Abstract

Motivated by the possible role of brain rhythms in cortical function, we postulate a cortical computation principle by which decoding 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. Syllabic segmentation utilizes flexible oscillators operating in the theta range (3–9 Hz) by tracking the input syllabic rhythms, and prosodic segmentation is driven by flexible oscillators in the delta range (0.5–3 Hz), tracking prosodic rhythms. A model (TEMPO) was developed which 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. The key properties that enable such accountability are: (i) the capability of the oscillators to track and stay locked to the input rhythm, and (ii) the cascaded nature of the oscillators within the array.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

sf298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

FinalReport_FA9550-11-1-0122.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

1. Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 3:238. doi:10.3389/fpsyg.2012.00238
2. Ghitza O., Giraud A-L and 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
3. Ghitza O. (2013). "The theta-syllable: a unit of speech information defined by cortical function." Front. Psychol. 4:138. doi: 10.3389/fpsyg.2013.00138
4. Doelling, K. B., Arnal, L. H., Ghitza, O. and Poeppel, D. (2014). "Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing." NeuroImage, 85:761–768. doi: 10.1016/j.neuroimage.2013.06.035
5. Ghitza O (2014) Behavioral evidence for the role of cortical theta oscillations in determining auditory channel capacity for speech. Front. Psychol. 5:652. doi:10.3389/fpsyg.2014.00652
6. 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
7. Fuglsang SA (2015) Towards predicting the intelligibility of time-compressed speech with silence gaps. Centre Applied Hearing Research, Technical University of Denmark. Master thesis. (Supervised by Ghitza O & Dau T).
8. Ghitza O (2016) Acoustic-driven delta rhythms as prosodic markers. Language, Cognition and Neuroscience. A special issue on "Brain oscillations in language comprehension and production". (Accepted.)

New discoveries, inventions, or patent disclosures:

Do you have any discoveries, inventions, or patent disclosures to report for this period?

No

Please describe and include any notable dates

Do you plan to pursue a claim for personal or organizational intellectual property?

Changes in research objectives (if any):

NO

Change in AFOSR Program Officer, if any:

Willard Larkin (Retired) to Patrick.Bradshaw

Extensions granted or milestones slipped, if any:

NO

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Sep 01, 2016 09:33:14 Success: Email Sent to: oghitza@bu.edu