Center for Biodynamics
Boston University

"Brain Rhythms in Speech Perception and Production"

Le Meridien Hotel, Cambridge, MA
November 8 - 9, 2008

Organizing Committee:
Oded Ghitza, Sensimetrics and Boston University (Chair)
Pierre Divenyi, EBIRE, Martinez, CA
Nancy Kopell, Boston University

Abstract

Speech is an inherently rhythmic phenomenon. Phonetic segments are articulated in syllabic "packages," which are spoken in cadence and reflect energy modulations with frequencies between 3 and 10 Hz. The intonation contour is also rhythmic, and this rhythmic aspect of speech is important for intelligibility and naturalness. Does speech rhythmicity reflect fundamental mechanisms internal to the brain? The purpose of this two-day workshop is to examine the role of endogenous brain oscillations in human speech perception and production, seeking to fill gaps in our models of speech perception and production and ultimately to improve the performance of systems for human-machine interaction, including automatic speech recognition and text-to-speech technology.

In speech perception there is a reasonable understanding of the function of the auditory periphery (i.e., the neural mechanisms responsible for generating a sensory representation), but little 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/psychophysical 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 to the primary auditory cortex, which allow us to design front-ends capable of representations that exhibit perceptually important speech information. On the other hand, recognition back-ends are based almost exclusively on statistical pattern recognition techniques, owing 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 only for tasks of limited complexity.

In speech production we have, at present, a fair understanding of functions related to the articulatory apparatus (e.g., the role of the lungs, larynx, vocal tract, and the nasal cavity in producing speech) but limited understanding of the identity and the nature of cortical mechanisms that transform abstract linguistic messages into the elaborate sequence of neuromotor commands associated with control of the vocal apparatus. While current text to speech (TTS) systems utilize back-ends based upon advanced models of articulation, the input to these models (phoneme sequences, duration of phonemes, intonation contour) is generated by front-ends governed by knowledge-based rules. Consequently, TTS systems are limited in their capability to translate text into naturally sounding, highly intelligible, high quality speech.

This workshop will examine the possible roles of brain rhythms in speech perception and production. Discussions will focus on (1) recent findings concerning the behavior of neural oscillations in animals and humans when performing tasks guided by sensory input, and (2) emerging computational models of rhythm generation and of the possible roles of neural rhythms in the processing of sensory input, and (3) the application of this research to models of speech perception and production, and their relevance to speech technology.

The workshop will bring together ca. 45 invited participants, experts in various aspects of speech and neuroscience. The workshop will comprise six regular sessions of 90 minutes each. A session will consist of two 20-minute presentations, each followed by 10 minutes of commentary by a discussant (who will be asked to offer a view on the connection between the presentation and speech science/technology). Sessions will conclude with a 30-minute general discussion. Each workshop participant will have an assigned role, either as a speaker or a discussant. The workshop will conclude with a final 90-minute discussion session.

Program (Each session is 90 min long)

Day 1 (Saturday, November 8)

0. Welcome and Introduction (10 minutes) 8:30

1. *Session 1: Rhythm in Speech, Music and the Brain* 8:45 – 10:15
   - 1.1. Rhythm in Speech and Music (Aniruddh Patel) – 20 minutes
   - 1.2. Rhythms in the Brain (Charles Schroeder) – 20 minutes
   - 1.3. Discussants (10 minutes each) – Neil Todd, Steve Greenberg

2. *Session 2: Brain Imaging, Speech and Rhythm* 10:30 – 12:00
   - 2.1. Acoustics, Speech and Rhythms (David Poeppel) – 20 min
   - 2.2. Language Comprehension and Rhythms (Peter Hagoort) – 20 min
   - 2.3. Discussants (10 min each) – Alain de Cheveigné, Mark Liberman

Lunch 12:00

3. *Session 3: Circuitry for Neural Computation* 1:00 – 2:30
   - 3.1. Neural Computation with Spikes (John Hopfield) – 20 min
   - 3.2. Neuronal Phase Lock Loops (Ehud Ahissar) – 20 min
   - 3.3. Discussants (10 min each) – Lloyd Watts, Guy Brown

4. *Session 4: Auditory Models* 2:45 – 4:15
   - 4.1. Modulation Maps (Torsten Dau) – 20 min
   - 4.2. Cortical Processing (Christoph Schreiner) – 20 min
   - 4.3. Discussants (10 min each) – Anne-Lise Giraud, Bertrand Delgutte

Day 2 (Sunday, November 9)

5. *Session 5: Temporal Sequencing in Perception and Production* 8:30 – 10:00
   - 5.1. Temporal Structure in Brain Signals (Matias Palva) – 20 min
   - 5.2. Neural Mechanisms of Sequence Generation (Michale Fee) – 20 min
   - 5.3. Discussants (10 min each) – Steve Levinson, Louis Goldstein

6. *Session 6: Rhythm in Speech Production and Recognition* 10:15 – 11:45
   - 6.1. Coupled oscillators in speech production (Elliot Saltzman) – 20 min
   - 6.2. Coupled oscillators in recognition (DeLiang Wang) – 20 min
   - 6.3. Discussants (10 min each) – Frank Guenther, Barak Pearlmutter

Lunch 11:45

7. *Session 7: Summary and General Discussion* 12:30 – 2:00

Closure 2:15

Rhythm in speech and music

Aniruddh D. Patel

The Neurosciences Institute
10640 John Jay Hopkins Dr.
San Diego, CA 92121.
apatel@nsi.edu

Rhythm is fundamental to speech and music. What do these two domains share in terms of rhythmic organization? There is a long history of interest in this topic by both linguists and musicologists, but remarkably few empirical explorations have been conducted. This is particularly striking since there are rich bodies of empirical research on rhythm within each domain.

In this presentation I suggest that progress in empirical comparative research depends on a clear distinction between periodic and nonperiodic rhythms in human auditory cognition. I will argue that speech and music have fundamental differences in terms of periodic rhythms, and important connections in terms of nonperiodic rhythms. Evidence for this argument draws on diverse strands of evidence, including quantitative comparisons of rhythmic patterns in speech and music, research on rhythm perception, and data from cognitive neuroscience.

Underlying this argument is a basic point about the definition of rhythm. "Rhythm" is a widely-used term in many fields (e.g., speech science, music cognition, neuroscience), and can mean different things to different people. Hence it is important to be explicit about what one means by rhythm. For many researchers, rhythm denotes periodicity, in other words, a pattern repeating regularly in time. I believe this definition is too narrow, and that a broader definition is needed that encompasses both periodic and nonperiodic rhythms. Hence I offer the following definition: Rhythm is the systematic patterning of timing, accent, and grouping in sequences of events. I welcome critiques of this definition (and concrete suggestions for alternatives) during the workshop.

Relevant readings
[First three available at http://www.nsi.edu/users/patel/publications.html]

Patel, A.D., Iversen, J.R., & Rosenberg, J.C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America,* 119:3034-3047.

Iversen, J.R., Patel, A.D., & Ohgushi, K. (in press). Perception of rhythmic grouping depends on auditory experience. *Journal of the Acoustical Society of America*.

Patel, A.D. (2006). Musical rhythm, linguistic rhythm, and human evolution. *Music Perception,* 24:99-104.

Ch. 3 (Rhythm) of Patel, A.D. (2008). *Music, Language, and the Brain*. NY: Oxford Univ. Press

Neuronal Oscillations and Visual Amplification of Vocal Communication

Charles E. Schroeder

Cognitive Neuroscience and Schizophrenia Program
Nathan Kline Institute
Department of Psychiatry
Columbia University College of Physicians and Surgeons
E-mail: Schrod@nki.rfmh.org

It is widely recognized that viewing a speaker's face enhances vocal communication, although the precise neural substrates of this phenomenon remain unknown. Drawing on work from a number of laboratories, we have proposed that the enhancement effect utilizes the ongoing oscillatory activity of local neuronal ensembles in primary auditory cortex. The idea that oscillations and oscillatory synchrony useful or even critical to brain operations has been debated extensively over the last decade. Recent evidence, however, lends weight to the hypothesis we advance here: *visual cues amplify auditory cortical processing of accompanying vocalizations by shifting the phase of ongoing neuronal oscillations so that the auditory inputs tend to arrive during a "high excitability state."* There are several facts about oscillations that combine under this idea. First, neuronal oscillations reflect rhythmic shifting of neuron ensembles between high and low excitability states. Second, oscillatory phase in auditory cortex clearly can be re-set by heteromodal (non-auditory) input. Third, cross-frequency oscillatory coupling gives rise to an extended oscillatory complex that is remarkably apt for representing the temporal sound energy patterns in human speech. Finally, attention strongly enforces the entrainment of cortical oscillations to visual, as well as auditory event streams. Because in A-V speech, visual cues precede accompanying vocalizations, oscillations are "predictively" modulated by visual input.

The phase-reset amplification mechanism we describe here should generalize beyond audiovisual communication. Across a wide range of real-world events, generally recognized as "biological motion," prominent non-auditory stimuli are generated prior to auditory stimulus onset because some visible action is required to produce a sound. For example, when we observe someone striking a nail with a hammer or running past us, the rhythmic temporal pattern of arm swinging or legs moving precedes and predicts the temporal pattern of hammer strike and footfall noises, particularly as the visual-auditory lag increases with distance. Visual cues often predict auditory events, and are thus in a position to modulate auditory perception.

It is of fundamental importance that the rhythms of the natural environment have a truly striking parallel in the rhythms of neuronal oscillation in the brain. The fact that the internal oscillations can be driven by external events, and can influence neuronal processing of the same events, reinforces the view that they are instrumental rather than incidental to sensory processing.

Acoustics, speech, and rhythms: cognitive neuroscience data

David Poeppel

Department of Linguistics, Department of Biology
University of Maryland College Park
College Park MD 20742
E-mail: david.poeppel@gmail.com

Speech signals reflect temporal regularities that are a consequence of the production apparatus. The biomechanical properties of the production system – the jaw, the articulators, etc. – are such that certain restricted temporal patterns emerge. These regularities occur over different time scales, with some types of speech information occurring in the range of 100-300 ms, other types of information over shorter time scales, in the range of 20-80 ms. While there are other phenomena in the time domain that exploit both shorter and longer time constants (for example gap detection, <5 ms; context effects, 1000 ms), we hypothesize that there are *two privileged temporal regimes in the analysis of spoken language*. Temporal modulation in the range of 100 to 300 ms is commensurate with information at the syllabic scale. Modulation over shorter time constants reflects sub-syllabic/featural changes in the speech signal. Perceptual analysis over these two time scales yields representations that link the input signal to the linguistic information that forms the basis for comprehension. Signal manipulations/degradations that affect the integrity of the information at these scales will compromise speech intelligibility.

We use techniques from cognitive neuroscience to investigate how the concurrent analysis of auditory information at multiple timescales is reflected in brain signals. Multi-time resolution analysis is an effective strategy employed by the visual system – as well as an many artificial systems. We extend this notion to auditory cognition. Both electrophysiological (MEG) and hemodynamic (fMRI) experiments are being conducted to test the hypothesis that information is being extracted on these two time scales.

Focusing on one of the time constants, I present data from human auditory cortex suggestive of a privileged role for processing at the 3-8 Hz (syllabic) rate (theta band). In particular, neurophysiological recordings using magnetoencephalography (MEG) demonstrate that oscillatory brain activity in the theta band correlates in systematic ways with the speech signal and with speech intelligibility. A typical SNR signal at normal speaking rates is associated with cortical theta band information (specifically phase). A view across speech and non-speech experiments suggests that information at the theta rate plays a foundational role for speech understanding, which in turn implicates the *syllable as an elementary processing unit*.

The mystery of language related brain oscillations

Peter Hagoort

Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behaviour
&
Max Planck Institute for Psycholinguistics
Peter.hagoort@fcdonders.ru.nl

Summary

Language comprehension involves two basic operations: the retrieval of lexical information (such as phonologic, syntactic, and semantic information) from long-term memory, and the unification of this information into a coherent representation of the overall utterance. Neuroimaging studies have provided detailed information on which areas of the brain are involved in these language-related memory and unification operations. However, much less is known about the dynamics of the brain's language network. Based on the current literature the following picture seems to emerge: memory retrieval operations are mostly accompanied by increased neuronal synchronization in the theta frequency range (4–7 Hz). Unification operations, in contrast, induce high-frequency neuronal synchronization in the beta (12–30 Hz) and gamma (above 30 Hz) frequency bands. However, this picture is far from clear, and the results are not always consistent across studies. Many conceptual, methodological and neurophysiological issues that are central to a better understanding of the phenomena under study remain to be solved. Two issues will be discussed in more detail:

1. Modulations of power and coherence have been found in at least 4 different frequency bands (theta, alpha, lower-beta and gamma). Is it possible to assign different functional roles (i.e., different aspects of language processing) to the different frequency bands?

2. In reviewing existing studies it is striking that, in contrast to the consistency of the results in terms of which frequency bands are affected by language processing, there seems to be an alarming inconsistency across studies in the topographic distribution of power and coherence changes (even within similar frequency bands and with similar experimental manipulations).

The 'Many Are Equal' neural algorithm and spike-timing based computation in speech processing

J. J. Hopfield

Carl Icahn Laboratory
Princeton University
Princeton NJ 08544
E-mail: hopfield@princeton.edu

Much of the effectiveness of the human brain in difficult tasks such as processing sound into words or separating scenes into multiple objects must come from the choice of what might be termed the 'neurobiological algorithms', approximate algorithms that are readily and collectively implemented by the 'hardware' and dynamics of neurobiology. The 'Many are Equal' (MAE) algorithm answers the question

> Given a large set of analog variables X_k is it true that there exists a substantial subset (the 'Many') X_n such that $X_n \approx X_{n'}$ for all n, n' in the subset?

This unlikely algorithm has multiple implementations in networks of spiking neurons, all involving action potential synchrony and collective rhythms. In some implementations the rhythms are caused by the MAE operation itself. In other implementations, a common collective rhythm (such as gamma or theta) developed by another set of neurons provides the mechanism of common synchronization. It is not yet known whether neurobiology actually makes use of this procedure, but we can examine it in the context of simulations and engineering.

I will explain and demonstrate two applications of this algorithm to very elementary examples of speech processing, for extraction of syllables or short words and for the recognition of shorter segments such as diphones. Each case involves a mapping of a biological representation of speech onto the variables of the MAE algorithm, and the implementation of the MAE procedure by a network of spiking neurons. In each case, the dynamics of the network can be seen to be related to engineering aspects of speech processing such as time warp and cepstral coefficients. In both cases, the intrinsic property of the MAE operation, in which by definition it ignores outlier data in the signal pattern being interpreted, provides useful interference rejection abilities.

Predictive decoding of the temporal envelope of speech with neuronal phase-locked loops

Ehud Ahissar

Department of Neurobiology
The Weizmann Institute
Rehovot, Israel
E-mail: ehud.ahissar@weizmann.ac.il

Listeners can adapt to 2-4-fold variations of syllable rate without losing comprehension. This could be achieved by syllable decoding mechanisms that are triggered by syllable onset. However, syllable onset often contains information that is critical for syllable decoding, which poses a serious problem for onset-triggered mechanisms. One solution for this problem is employing a predictive decoding mechanism, which predicts onset time. Such a mechanism can be implemented by neuronal phase-locked loops (NPLLs). In analogy to electronic PLL, NPLL is composed of an intrinsic rate-controlled-oscillator (RCO) which establishes a negative closed loop with a phase detector (PD), whose other input is the speech signal. If tuned correctly, and if the envelope frequency (i.e., syllable rate) is within the working range of the loop, the RCO will track the envelope's frequency and in fact predict the onset of each syllable. If the output of the NPLL triggers syllable-processing circuits, syllable onsets will not be lost.

While the operation of a PLL-like mechanism in speech processing had not yet been demonstrated, there is a significant body of evidence that allows the operation of such a mechanism in the brain. Specifically, NPLLs can be implemented by thalamocortical loops, in which temporal comparison takes place in the thalamus, most likely non-lemniscal nuclei. With thalamocortical NPLLs, the signal indicating timing differences, produced by the non-lemniscal thalamus, is fed back to the cortex where it is used to update the frequency of the intrinsic oscillators. It is the intrinsic "clock" that sets the pace for segmentation, and not the speech signal. The rate of the speech signal only updates the intrinsic clock, and makes it a better predictor of the following input rate. A primacy of the intrinsic rhythm in envelope decoding is consistent with the observation that the psychological moment of occurrence (the P-center) of a syllable is based on a comparison of the speech signal with some kind of internal "temporal ruler" rather than on the speech signal per se.

An NPLL envelope decoder predicts that the ability of listeners to adapt to varying speech rates depends on the dynamic range of their cortical oscillations. Indeed, we have demonstrated a noticeable correlation between the modal frequency of cortical oscillations in humans, recorded via MEG, and their comprehension thresholds for accelerated speech. Further experimental testing should determine to what extent, and with which dynamics, cortical oscillations follow changes in speech rate, whether cortical dynamic ranges can be increased by training, and whether such increases will facilitate comprehension of varying-rate speech.

Spectro-temporal processing of complex sounds in the human auditory system

Torsten Dau

Department of Electrical Engineering
Technical University of Denmark
DK-2800 Lyngby
E-mail: tda@elektro.dtu.dk

The perception of complex sounds like speech is critically dependent on the faithful representation of the signal's spectral and temporal modulations in the auditory system. Several stages of auditory processing are considered to be crucial for a robust representation of such spectro-temporal modulations and a deficiency in any of these processing stages is likely to result in a deterioration of the entire system's performance.

This presentation considers models of auditory processing and *perception* of modulated sounds. The modelling is inspired by neurophysiological findings but reflects an "effective" modelling strategy that does not allow conclusions about details of signal processing at a neuronal level. On the other hand, since the effective model accounts for a large variety of perceptual data, such as spectro-temporal masking patterns and speech intelligibility results, this suggests certain processing principles which in turn motivate the search for neural circuits in corresponding physiological studies. The modelling assumes as one of the key elements an amplitude modulation filterbank at the output of each cochlear filter (Dau *et al*., 1997; Jepsen *et al*., 2008). The modulation filterbank realizes a limited-resolution decomposition of the temporal modulations whereby the parameters of the filterbank are not directly related to the parameters from physiological models that describe the transformation from a temporal neural code into a rate-based representation of AM in the auditory brainstem and cortex (e.g., Nelson and Carney, 2004). The output of the preprocessing, i.e., the "internal representation" of the acoustical input signal, has been used in a variety of applications, e.g., for assessing speech quality, predicting speech intelligibility and as a front-end for automatic speech recognition.

A conceptually similar approach has been presented by Shamma and co-workers (e.g., Elhilali *et al.*, 2003; Chi *et al.*, 2005). They described a model that includes an additional "dimension" in the signal analysis. They suggested a spectrotemporal analysis of the envelope, motivated by neurophysiological findings in the auditory cortex (e.g., Schreiner and Calhoun, 1995). In their model, a "spectral" modulation filterbank is combined with the temporal modulation analysis, resulting in two-dimensional spectro-temporal filters. Thus, in contrast to the implementation presented above, their model contains joint (and inseparable) spectro-temporal modulations. In conditions where both temporal and spectral features of the input are manipulated, the two models respond differently. The model of Shamma and co-workers has been utilized to account for spectro-temporal modulation transfer functions, the assessment of speech intelligibility as well as for the prediction of musical timbre.

References:

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

Chi, T., Ru, P., Shamma S.A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**, 887-906.

Jepsen, M.L., Ewert, S.D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**, 422-438.

Elhilali, M., Chi, T., and Shamma, S. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech. Commun. **41**, 331–348.

Nelson, P. C., and Carney, L. H. (**2004**). "A phenomenological model of peripheral and central neural responses to amplitude-modulated tones," J. Acoust. Soc. Am. **116**, 2173–2186.

Schreiner, C. E., and Calhoun, B. (**1995**). "Spectral envelope coding in cat primary auditory cortex: Properties of ripple transfer functions," Aud. Neurosci. **1**, 39–61.

Temporal Modulation Processing in Auditory Cortex

Christoph E. Schreiner

W.M. Keck Center for Integrative Neuroscience
Dept. of Otolaryngology-HNS
University of California, San Francisco
San Francisco, CA 94143
E-mail: chris@phy.ucsf.edu

Auditory cortical neurons show several seemingly disparate properties when considering the encoding of temporal stimulus information. Individual acoustical events can be marked by cortical activity with high precision in the millisecond range. Repetitive events, however, are commonly only marked if they occur less then 20-40 times per second. Stimuli containing higher repetition rates can be encoded by a rate code. Finally, intrinsic temporal response oscillations can interact with external event coding. We will discuss several issues surrounding auditory cortical temporal coding.

1) Modulation Encoding
   The encoding of temporal information depends critically on the response mode of neurons, most notably whether neurons are tonic or phasic responders. Consequences for rate versus time codes will be discussed in particular with regard to studies in awake versus anesthetized preparations.
2) Modulation Representation
   The range of temporal information represented by different neurons depends on the type of neuron (e.g. excitatory versus inhibitory) and the expressed response mode. Additional factors to be considered when discussing temporal information transfer are the cortical layer and the cortical area, such as in the core versus belt versus parabelt classification of auditory cortical fields. Within each field the spatial distribution of the temporal code may vary. The characterization of temporal response properties often uses spike-triggered averaging to extract the preferred temporal response ranges and define a temporal filter (and response nonlinearity). Recent findings indicate that a single filter is not sufficient to full characterize the cortical response behavior but that at least a second, independent filter is required for a more complete response description.
3) Modulation Filter Generation and Plasticity
   The generation of cortical temporal response preferences is largely influenced by cortico-cortical processes and not solely determined by thalamo-cortical feed-forward mechanisms. Plasticity of temporal properties due to interference during cortical development and through behaviorally relevant experience during adulthood can dramatically shape the processing capacity for temporal information.

Temporal Structure in Brain Signals

Matias Palva

Neuroscience Center
University of Helsinki
PO Box 56, 00014 Univ. Helsinki, Finland
E-mail: matias.palva@helsinki.fi

Electro- and magnetoencephalography (EEG and MEG) measure human brain activity non-invasively and with millisecond-range temporal resolution. EEG and MEG signals are produced by synchronous post-synaptic currents in large neuronal populations (>10^5–10^6 neurons). Hence, the mere existence of EEG and MEG signals indicates that there are temporal structures in the underlying brain activity. The 1/*f*-type power spectrum of the EEG shows that these temporal structures are power-law distributed in a very wide (0.01–100 Hz) frequency range[1]. In addition, amplitude fluctuations in specific narrow frequency bands also have a 1/*f*-type power spectrum[2]. Brain activity thus appears to be temporally and spectrally scale-free and power-law distributed. This is evident in both perception and action. For instance, the detection performance of very weak somatosensory stimuli varies with scale-free dynamics and is biased both by very slow (0.01-0.1 Hz) brain waves and by the amplitude fluctuations of much faster (> 1 Hz) neuronal oscillations[1]. Many features of action production, such as reaction times and error rates also fluctuate over time in a scale-free fashion.

The "scale-free" image of brain activity is, however, complemented by many observations of frequency-band-limited, *i.e.*, "scale-specific", task effects in EEG and MEG experiments[3]. Oscillations in different frequency bands are likely to have distinct neuronal generators, synchronization properties, and consequently distinct functional specializations. Phase synchronization within a frequency band is generally accepted to be a means for the communication and integration between anatomically distributed brain areas. What are the mechanisms that mediate the communication, integration, and coordination of neuronal assemblies in distinct frequency bands? Two candidate mechanisms have been identified so far: nested oscillations[4] and *n:m*-phase synchrony[5]. In nested oscillations, the amplitude of a fast oscillation is modulated by the phase of a slow oscillation. In *n:m*-phase synchrony, the phase of a fast oscillation is locked to the phase of the slow oscillation. These cross-frequency interactions span the entire frequency range of brain activities and bind spectrally distributed assemblies into multi-band oscillatory hierarchies[4,5].

In my presentation, I highlight power-law scaling behavior and within-/cross-frequency-phase interactions as temporal structures that are relevant for perception and action production.

---

[1] Monto et al., J Neurosci, 2008, 28:8268–8272.
[2] Linkenkaer-Hansen et al., J Neurosci, 2001, 21:1370–1377.
[3] Palva et al., J Neurosci, 2005, 25:5248–5258.
[4] Lakatos et al., J Neurophysiol, 2005, 94:1904-1911.
[5] Palva et al., J Neurosci, 2005, 25:3962–3972.

Mechanisms of Sequence Generation in Vocal Production

Michale S. Fee

Department of Brain and Cognitive Sciences
McGovern Institute for Brain Research
Massachusetts Institute of Technology
Cambridge, MA 02139
E-mail: fee@mit.edu

From the most basic motor behaviors, such as locomotion, to the most complex, such as speech and language, the timing and serial ordering of movements are crucial. For some simple oscillatory behaviors, in which the movement evolves on a single timescale, it has been possible to identify the particular neurons and biophysics that control the temporal dynamics of the behavior — for example, pacemaker neurons in the stomatogastric ganglion, or the oscillator network that controls swimming in the leech. But what mechanisms underlie more complex learned behaviors that have structure on many timescales? Birdsong exhibits a remarkably precise and hierarchically organized temporal structure mediated by a number of distinct, well-studied motor nuclei, which allows for an unprecedented view into the neuronal mechanisms of sequence generation. Adult zebra finches generate a stereotyped sequence of sounds with structure at several timescales, from 10ms to 100ms to 1sec.

Two vocal-related brain regions have been implicated in the control of the temporal structure of bird song: HVC and RA. HVC projects to RA, which in turn projects to the vocal motor neurons as well as midbrain vocal control and brainstem respiratory areas. Neurons in HVC burst extremely sparsely during singing, each generating a single brief (~6 ms) burst of spikes at a particular moment in every repetition of the song. In addition, different HVC neurons burst at different times throughout the song. Thus, HVC neurons appear to code for time, or temporal order, in the song sequence.

But where are the dynamics that control the timing of these brief events in HVC? Are these bursts are driven by timing circuitry in other brain areas that project to HVC? Or do these HVC neurons, as a population, generate a wave of activity that propagates through HVC — like a chain of dominoes — which then controls the timing of the song? We have developed a new technique for localizing temporal dynamics within brain circuitry, taking advantage of the fact that the speed of brain processes is strongly temperature dependent. If the circuitry in a particular brain area is involved in controlling song timing, then localized mild cooling of that area should slow the song.

Remarkably, we find that cooling of nucleus HVC results in a slowing of song timing at all timescales. In contrast, cooling RA has no effect on song timing. Our results suggest that vocal timing in the songbird is controlled by a chain of activity, possibly largely mediated by chain-like synaptic connectivity within HVC. We have used local manipulation of brain temperature to identify components within the avian song system that control the timing of a complex behavioral sequence. A similar approach may be broadly useful to localize specialized brain circuits that control the timing of other behaviors, and other forms of brain dynamics.

Coupled oscillators in speech production

Elliot Saltzman

Boston University
Boston, MA 02215
&
Haskins Laboratories
New Haven, CT 06511
E-mail: esaltz@bu.edu

The original task-dynamic model of speech production incorporated the theoretical tenets of Articulatory Phonology and provided a dynamics of inter-articulator coordination for single and co-produced constriction gestures, given a gestural score that specifies a time-dependent vector of gestural activations for a given utterance. More recently, the model has been significantly extended to provide a framework for investigating the higher order dynamics of prosodic phrasing, syllable structure, lexical stress, and the prominence (accentual) properties associated with higher level prosodic constituents (e.g., foot, word, phrase, sentence). There are two new components in the model. The first is an ensemble of *gestural planning oscillators* that defines a dynamics of gestural score formation in that, once the ensemble reaches an entrained steady-state of relative phasing, the waveform of each oscillator is used to trigger the activation function of that oscillator's associated constriction gestures. The second component is a set of *modulation gestures* (μ-gestures) that, rather than activating constriction formation and release gestures in the vocal tract, serve to modulate the temporal and spatial properties of all concurrently active constriction gestures. Modulation gestures are of two types: temporal modulation gestures (μ_T-gestures) that alter the rate of utterance timeflow by smoothly changing all frequency parameters of the planning oscillator ensemble; and spatial modulation gestures (μ_S-gestures) that spatially strengthen or reduce the motions of constriction gestures by smoothly changing the spatial target parameters of these constriction gestures. Key to the representation of prosodic phrasing has been use of clock-slowing temporal modulation gestures (called prosodic gestures [π-gestures] in previous work) that are locally active in the region of phrasal boundaries, and that slow the rate of utterance timeflow in direct proportion to the strength of the associated boundary. Central to the representation of syllable structure is the use of a *coupling graph* that defines the existence and strength of coupling in the network of gestural planning oscillators. Concepts from graph theory have been crucial to understanding how hypothesized differences among coupling graphs have correctly predicted empirically demonstrated intra-syllablic differences between onsets and codas in both the mean values and variabilities of C-C, C-V, and V-C timing patterns. In this talk, I will describe recent developments to the task-dynamic toolkit (original task-dynamic model, planning oscillator ensemble, and modulation gestures) and how they have been used to interpret and simulate experimental data on the interactions of stress and prominence in shaping the kinematic details of speech production.

Speech Segregation by Oscillatory Correlation

DeLiang Wang

Department of Computer Science and Engineering, and Centre for Cognitive Science
The Ohio State University
Columbus, Ohio 43210
E-mail: dwang@cse.ohio-state.edu

What neural mechanisms underlie auditory scene analysis? Both theoretical and empirical investigations of the brain point to the mechanism of oscillatory correlation as a plausible paradigm for scene representation. In this presentation I describe an oscillatory correlation approach to the problem of speech segregation, or cocktail-party processing. In the oscillatory correlation approach, a perceptual stream corresponds to a synchronized assembly of neural oscillators and different streams correspond to desynchronized oscillator assemblies. This approach has been employed for double-vowel separation and segregation of voiced speech. An oscillator model for double-vowel separation synchronizes auditory channels that define the spectral components of each vowel on the basis of periodicity analysis. This model is able to replicate the perceptual observation that listeners' ability to identify concurrent vowels improves with increasing difference in fundamental frequency between the vowels. For voiced speech segregation, a two-layer network of relaxation oscillators is used. The first layer performs the task of auditory segmentation whereby an auditory scene is broken into a collection of auditory segments, each of which corresponds to a contiguous time-frequency region. The second layer performs the task of grouping, in which segments are organized into distinct streams. Lateral connections between oscillators encode harmonicity and proximity in frequency and time. Prior to the oscillator network are a model of the auditory periphery and a stage in which mid-level auditory representations, such as correlogram, are formed. This model of speech segregation has been evaluated using a corpus of voiced speech mixed with a variety of interfering sounds. Further developments of this model are discussed. Finally, I will speculate on possible roles of oscillatory correlation in the broad framework of computational audition.