UNIVERSITY OF CALIFORNIA,
IRVINE

# Cortical Entrainment to Speech: Effects of Attention, Intelligibility, and Regularity

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Lucas Samuel Baltzell

Dissertation Committee:
Professor Virginia M. Richards, Co-Chair
Professor Ramesh Srinivasan, Co-Chair
Professor Fan-Gang Zeng

2018

# DEDICATION

To family and friends,
and to all those in whom you find better versions of yourself

Such is the course of deeds that move the wheels of the world:
small hands do them because they must,
while the eyes of the great are elsewhere

J.R.R. Tolkien
Lord of the Rings

The truth will set you free,
but not until it is finished with you

David Foster Wallace
Infinite Jest

# TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I would like to thank my committee members en masse, who at varying times and to varying degrees have provided me with the ability and opportunity to achieve any success that I may have achieved.

I would specifically like to thank my committee co-chair, Professor Virginia M. Richards, who by sheer force of repetition managed to teach me the basics of psychoacoustic methods, and who by any measure, was the most helpful and attentive advisor as could be hoped for.

I would specifically like to thank my committee co-chair, Professor Ramesh Srinivasan, whose blend of intellectual honesty and practical realism helped me navigate the psychological pressures of pursuing a career in research.

I would like to thank Jon Venezia, for being a clear thinker and a good man.

Financial support for the work presented in this dissertation was provided by laboratory grants/funds and by the Center for Hearing Research.

# CURRICULUM VITAE

Department of Cognitive Sciences
University of California, Irvine
184 Social Science Laboratory
Irvine, CA 92697-5100

Luke Baltzell
lucassbaltzell@gmail.com
Hearing Lab

## EDUCATION

| | | |
|---|---|---|
| 2013-2018 | University of California, Irvine<br>*Ph.D Psychology* | Irvine, CA |
| 2013-2016 | University of California, Irvine<br>*M.S. Cognitive Science* | Irvine, CA |
| 2006-2010 | Reed College<br>*B.A. Psychology* | Portland, OR |

## AWARDS

| | | |
|---|---|---|
| 2016 | Outstanding Graduate Student<br>Center for Hearing Research, University of California, Irvine | Irvine, CA |
| 2009-2010 | Commendation for Excellence in Scholarship<br>Administration Committee of Reed College | Portland, OR |
| 2008-2009 | Commendation for Excellence in Scholarship<br>Administration Committee of Reed College | Portland, OR |
| 2007-2008 | Commendation for Excellence in Scholarship<br>Administration Committee of Reed College | Portland, OR |

## GRANTS & FELLOWSHIPS

| | | |
|---|---|---|
| 2014-2016 | Interdisciplinary Training Program in Hearing Research, Center for Hearing Research (NIH T32 DC010775)<br>University of California, Irvine | Irvine, CA |
| 2010 | Initiative Grant for Undergraduate Research<br>Office of the President | Portland, OR |

## PUBLICATIONS: *Peer-Reviewed*

**Baltzell, LS.**, Srinivasan, R., and Richards, VM. (2017). The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *Journal of Neurophysiology*, 118: 3144–3151.

**Baltzell, LS.**, Horton, C., Yi, S., Richards, VR., D'Zmura, M., and Srinivasan, R. (2016) Attention selectively modulates cortical entrainment in different region of the speech spectrum. *Brain Research*, 1644: 203-212.

Billings, CJ., Penman, TM., Ellis, EM., **Baltzell, LS.**, and McMillan, GP. (2016) Phoneme and Word Scoring in Speech-in-Noise Audiometry. *American Journal of Audiology*, 25: 75-83.

Papesh, MA., Billings, CJ., and **Baltzell LS**. (2014). Background noise can enhance cortical auditory evoked potentials under certain conditions. *Clinical Neurophysiology*, doi:10.1016/j.clinph.2014.10.017

**Baltzell, LS.** and Billings, CJ. (2013). Sensitivity of offset and onset cortical auditory evoked potentials to signals in noise. *Clinical Neurophysiology*, 125: 370-380.

Billings, CJ., Papesh, MA., Penman, TM., **Baltzell, LS.**, and Gallun, FJ. (2012). Clinical Use of Aided Cortical Auditory Evoked Potentials as a Measure of Physiological Detection or Physiological Discrimination. *International Journal of Otolaryngology*, doi:10.1155/2012/365752

PUBLICATIONS: *Other*

**Baltzell, Lucas S.** *Interaction Between Syntax Processing in Language and Music as a Function of Bilingualism: An ERP Study*. Diss. Reed College, 2010. OCLC #: 624366709.

CONFERENCE POSTERS

**Baltzell, LS.**, Xia, J., Kalluri, S. (2018) "Efficient characterization of compression ratio preference in hearing-impaired listeners." 41th ARO Mid-Winter Meeting; February 11-15.

**Baltzell, LS.**, Richards, VM., Srinivasan, R. (2017) "The cortical entrainment response to speech shows no effect of stimulus intensity," 40th ARO Mid-Winter Meeting; February 11-15.

**Baltzell, LS.**, Richards, VM., Srinivasan, R. (2016) "Exploring the effects of intelligibility on cortical entrainment," Neuroscience 2016; November 12-16.

**Baltzell, LS.**, Srinivasan, R., Richards, VM. (2016) "Effects of task demand and intelligibility on the cortical entrainment response," 8th Annual Society for the Neurobiology of Language Conference; August 17-20.

**Baltzell, LS.**, Horton, C., Shen, Y., Richards, VM., and Srinivasan, R. (2015) "Attentional filtering through cortical entrainment shows sensitivity to frequency," 38th ARO Mid-Winter Meeting; February 16-20.

**Baltzell, LS.** and Billings, CJ. (2013) "Neural Encoding of Speech in Noise: Comparing Auditory Evoked Offset and Onset Responses," 36th ARO Mid-Winter Meeting; February 16-20.

**Baltzell, LS.**, Dillman, G., Gallun, FJ., Molis, MR., Konrad-Martin, D., Billings, CJ. (2012) "Auditory Brainstem Encoding of Envelope and Fine Structure: Recording the Frequency Following Response," Northwest Auditory & Vestibular Meeting; October 25-26.

Papesh, MA., **Baltzell, LS.**, Billings, CJ. (2012) "Can Background Noise Enhance Cortical Encoding?," Northwest Auditory & Vestibular Meeting; October 25-26.

## INVITED TALKS

"The role of cortical entrainment in speech perception: some considerations" (January, 2018). *Hearing Research Center Seminar Series*, Boston University

"Attentional filtering through cortical entrainment shows sensitivity to frequency" (May, 2016). *Eleventh Annual Center for Hearing Research Symposium*, University of California, Irvine

## PROFESSIONAL SERVICE

- Ad Hoc Reviewer, *JASA*
- Ad Hoc Reviewer, *Hearing Research*
- Ad Hoc Reviewer, *Ear and Hearing*

## PROFESSIONAL AFFILIATIONS

- Member, *Association for Research in Otolarynology*

## WORK EXPERIENCE

2017 Starkey Hearing Research Center Berkeley, CA
**Research Intern**

2017 – 2018 University of California Irvine Irvine, CA
**Teaching Assistant**

2013 – 2014 University of California Irvine Irvine, CA
**Teaching Assistant**

2011 – 2013 National Center for Rehabilitative Auditory Research Portland, OR
**Research Assistant** for Dr. Curtis Billings

2011 Department of Linguistics, UCLA Los Angeles, CA
**Research Assistant** for Dr. Carson Schutze

2010 – 2011 SUN After-School Program, Grout Elementary Portland, OR
**Instructor** through Schools Uniting Neighborhoods (SUN)

2009 Department of Psychology, Reed College Portland, OR
**Research Assistant** for Dr. Enriqueta Canseco-Gonzalez

2009 Department of Psychology, Reed College Portland, OR
**Teaching Assistant**

REFERENCES

Dr. Virginia Richards – Professor of Cognitive Science, University of California, Irvine
949.824.2051
v.m.richards@uci.edu

Dr. Ramesh Srinivasan – Professor of Cognitive Science, University of California, Irvine
949.824.2696
r.srinivasan@uci.edu

Dr. Fan-Gang Zeng – Professor of Otolaryngology, University of California, Irvine
949.824.1539
fzeng@uci.edu

Dr. Sridhar Kalluri – Director of Starkey Hearing Research Center, Starkey Hearing Technologies
Sridhar_Kalluri@starkey.com

Dr. Curtis J. Billings – Research Investigator, National Center for Rehabilitative Auditory Research (NCRAR)
503.220-8262 x54574
Curtis.Billings2@va.gov

# ABSTRACT OF THE DISSERTATION

Cortical Entrainment to Speech: Effects of Attention, Intelligibility, and Regularity

By

Lucas Samuel Baltzell

Doctor of Philosophy in Psychology

University of California, Irvine, 2018

Professor Virginia M. Richards, Co-Chair

Professor Ramesh Srinivasan, Co-Chair

A neural response can be recorded from the scalp that follows ongoing fluctuations in speech energy. This response has been labelled cortical “entrainment”, and while a number of theoretical proposals have been offered regarding the functional role of cortical entrainment in speech perception, the degree to which cortical entrainment reflects peripheral vs. central representations of speech, and acoustic vs. linguistic processes remains unclear. First, we show that the entrainment response reflects differential attentional weights applied across auditory frequency channels, suggesting the entrainment response follows fluctuations in speech within peripheral frequency channels. Second, we show that the strength of the entrainment response to natural fluctuations in speech energy does not depend on whether or not the speech is intelligible, suggesting that the response reflects primarily acoustic rather than linguistic processes. Third, we show that with unnaturally periodic speech stimuli, an entrainment response follows changes in hierarchically-organized levels of linguistic information, independent of acoustic energy, suggesting that under certain conditions, the entrainment response reflects linguistic processes. However, we observe the same entrainment response to hierarchically-organized levels of musical information, suggesting that the cortical processes underlying this entrainment response are not necessarily specific to speech. Together, these results suggest that cortical entrainment can reflect both peripheral and central processes, and that these processes may not be unique to speech.

# INTRODUCTION

Speech is a complex acoustic signal that contains fluctuations in energy and linguistic information at multiple, hierarchically organized timescales (Friederici, 2002; Mattys et al., 2005; Ding et al., 2017). While we know that information is extracted at these different timescales, and that this information is passed through different cortical structures resulting in meaning extraction (Hickok & Poeppel, 2007), the neural mechanisms supporting this hierarchical processing are not fully understood.

In the last decade, research investigating these neural mechanisms has flourished, driven by the proposal that cortical oscillations entrain to different levels of the speech hierarchy. Specifically, cortical entrainment is thought to support the parsing of speech by aligning hierarchically-coupled oscillations to the natural fluctuations of the speech stimulus, effectively packaging the acoustic input into discrete, hierarchically organized units optimized for linguistic processing (Ghitza, 2011; Giraud & Poeppel, 2012). The evidence for this entrainment proposal can be grouped into two categories: near-field mechanistic, and far-field observational.

The near-field mechanistic evidence concerns the functional role of cortical oscillations. Synchronous oscillations in the cortex arise naturally from both synaptic delays in local neural circuits and transmission delays in long-range neural connections (Nunez & Srinivasan, 2006). For certain local neural circuits, it has been demonstrated that oscillatory activity reflects synchronized modulations of neuronal excitability (Sanchez-Vives & McCormick, 2000; Buzsaki, 2002; Buzsaki & Draguhn, 2004). Based in large part on these findings, and on findings that attention can lead to increased neural synchrony (e.g. Fries et al., 2001), it has been suggested that oscillations enable coordination at different timescales across and within cortical networks, and in this way govern perceptual and cognitive processes (e.g. Fries, 2005). Lending support to this claim, recent studies have shown that cortical oscillations can hierarchically couple such that the phase of lower-frequency oscillations can modulate the power at higher-frequency oscillations, that this coupling is task-specific, and that behavioral performance depends on the phase of these oscillations (Lakatos et al., 2005; Canolty et al., 2006; Lakatos et al., 2008; Lakatos et al., 2013). These studies also show that the phase of these oscillations can be reset by transient responses to sound onsets[1]. The near-field mechanistic evidence suggests then, that oscillations in the cortex can reflect fluctuations in neuronal excitability, and that perception can depend on the phase of cortical oscillations.

The far-field observational evidence concerns recordings from large synchronous populations of neurons (EEG, MEG) that correlate with fluctuations in the speech stimulus. For instance, there is a robust, attentionally-dependent response that correlates with the low-pass filtered stimulus envelope (Ding & Simon, 2012; Horton et al., 2013). While it is tempting to interpret this correlation as reflecting a phase-resetting oscillation that entrains to ongoing fluctuations in speech energy, such an interpretation is not necessarily warranted, as the cortical processes responsible for this correlation aren't necessarily known. For instance, speech can be modelled as a series of abrupt acoustic transients, and since cortical responses are evoked by acoustic transients, a cortical response will emerge that directly follows the acoustic envelope, whether or not these transients induce a phase-

[1] The phase-resetting property of these oscillations raises an interesting question about their purported role in speech perception. If we consider speech as a series of acoustic transients that generate a corresponding series of phase-resets, then perhaps fluctuations in neuronal excitability are single-cycle phenomena, designed to discretize a chunk of acoustic material following an abrupt onset.

reset of an ongoing oscillation of neuronal excitability. Indeed, replacing natural bursts of speech energy with sharper acoustic transients leads to a greater correlation between the acoustic envelope and the cortical response (Doelling et al. 2014). This result is noteworthy for two reasons. First, it suggests that auditory onset responses, which are known to be proportional to the sharpness of the sound onset, may be largely (if not completely) responsible for generating correlations between the cortical response and the speech envelope. Second, it suggests that the linguistic content of the speech signal does not contribute to the cortical response. While a number of studies have made claims to the contrary, design confounds have made these studies difficult to interpret (see Chapter 2), and when acoustic confounds have been removed, the addition of linguistic information does not seem to have an effect on the strength of the cortical response to the stimulus envelope.[2]

The apparent insensitivity of the cortical response to linguistic content may be partially due to the generally poor correlations between stimulus energy and perceived linguistic boundaries (Klatt, 1980).[3] To this end, Di Liberto et al. (2015) found that the cortical response to speech is best predicted using both acoustic-level and phoneme-level representations of the stimulus. This suggests that while the cortical response to the

[2] The entrainment proposal however, does not necessarily predict a dependence on linguistic content. It could be the case that a cortical oscillation entrains to the acoustic energy, parses the input into discrete syllabic chunks, and that in the absence of linguistic features, hierarchically-coupled oscillations simply have no further information to extract (Doelling et al., 2014). Given the perceptual interdependence of different levels of the linguistic hierarchy though (Mattys et al., 2005), a structure of cortical oscillations that are sensitive to linguistic content should be more powerful than a structure that is entirely dependent on acoustic energy.

[3] These poor correlations are potentially problematic for the entrainment proposal. If linguistic boundaries are not highly correlated with acoustic energy, why should oscillations entrain to the envelope? In other words, to the extent that acoustic boundaries and linguistic boundaries are uncorrelated, oscillations phase-locked to the acoustic envelope are as likely to be out-of-phase with linguistic units as they are likely to be in-phase.

envelope may be acoustically driven, the cortical response to speech is sensitive to purely linguistic changes.

Relatedly, Ding et al. (2016) presented single words at a rate of 4 Hz, with phrases constituting 2-word units (e.g. "smart dogs"), and sentences constituting 4-word units ("smart dogs dig holes"). The identified a cortical response that follows phrasal (2 Hz) and sentential (1 Hz) periodicities, distinct from and in addition to a cortical response that follows the 4 Hz periodicity of the stimulus envelope, neatly demonstrating the existence of cortical responses that follow fluctuations in linguistic information. However, while this study elegantly differentiates between entrainment to linguistic and acoustic cues, the speech stimuli used to make this distinction were artificially periodic, with modulation rates in natural speech being far less regular. It is therefore possible that cortical processes not typically involved in speech perception are being recruited.

While both Di Liberto et al. (2015) and Ding et al. (2016) have demonstrated cortical responses to fluctuations in linguistic information, it is not clear whether or not phase-resetting cortical oscillations are involved. If we suppose that information is passed from one cortical locus of linguistic processing to another, and that a signature of this transfer exists, a cortical response will be generated that appears to follow fluctuations in linguistic information. Such an explanation does not require reference to entraining oscillations of neuronal excitability, nor does it necessarily suggest coordination across multiple timescales. Neither however, does it exclude these explanations, and while future research may bridge the gap between near-field mechanistic explanations and far-field observational results, at present, the suggestion that speech perception is supported by hierarchically-coupled oscillations in neuronal excitability remains to be directly confirmed.

This agnosticism, unfortunately, will not be necessarily be mitigated by the experiments reported in this dissertation. Putting aside the question of whether or not phase-retting cortical oscillations are responsible though, it is simply the case that fluctuations in speech energy and speech information can generate a cortical response that is correlated with these fluctuations, and that these correlations can offer useful insights into the cortical processing of speech.

In chapter 1, we present evidence suggesting that the entrainment response follows fluctuations in speech energy within individual frequency channels of the auditory system, and that this response contributes to or reflects the differential allocation of attentional weights across these channels. In chapter 2, we show that when acoustic confounds are controlled for, the entrainment response to the speech envelope does not depend on the linguistic content of the stimulus. We also show that the strength of this envelope-following response is highly modulated by task demands, even in the absence of a competing stimulus. This underscores the previous findings that attention plays a critical role in modulating the strength of the entrainment response to the speech envelope, and suggests that task demands must be carefully controlled in future studies investigating this response. In chapter 3, we reproduce a key finding that an entrainment response can follow hierarchically-organized periodic fluctuations in linguistic information, and show that this response is not domain-specific to speech. Specifically, we show that hierarchically-organized musical stimuli can elicit a similar response pattern, suggesting that cortical processes recruited for the hierarchical processing of speech may also be recruited for the hierarchical processing of music.

### *A Note on Terminology*

In the chapters that follow, cortical responses that follow fluctuations sound energy/information will be referred to as “entrainment” responses. Keeping in mind the mechanistic qualifications offered above, the term cortical “entrainment” is not meant imply anything beyond how it is measured, as a correlation between two time series. In this sense, cortical entrainment does not necessarily refer to a single process, and the interpretation of a particular entrainment response should be grounded by the context in which it is elicited.

# CHAPTER 1

## INTRODUCTION

A number of studies have been published examining the effects of attention on neural responses that appear to track the temporal envelope of speech (and non-speech) in the auditory cortex (Kerlin et al., 2010; Ding & Simon, 2012a,b; Mesgarani & Chang, 2012; Ng et al., 2012; Power et. al., 2012; Horton et al., 2013; Horton et al., 2014; Zion-Golumbic et al., 2013; Ding & Simon, 2014; Ding et al., 2014; O'Sullivan, 2014; Di Liberto et al., 2015). This phenomenon is often referred to as cortical "entrainment," and while the underlying mechanisms are still unclear, it is thought to reflect important aspects of temporal processing (for a review, see Ding & Simon, 2014). Pointing to the correspondence between the modulation spectrum of speech and the power spectrum of cortical oscillations, it has been suggested that these oscillations play an active role in parsing the acoustic speech stimulus into discrete syllable-length units for linguistic processing (Ghitza, 2011; Giraud & Poeppel, 2012; Doelling et al., 2014). While this functional claim remains somewhat controversial, it appears clear that the envelope-tracking response reflects attentional mechanisms that enhance the response to the target and suppress the response to the distractor in a complex auditory scene (Ding & Simon, 2012a,b; Mesgarani & Chang, 2012; Horton et al., 2013).

It has been suggested that attention is modulating the envelope-tracking response at the level of the auditory object (e.g. Ding & Simon, 2012a,b; Ding & Simon, 2014). In other words, attention is being applied to some neural reconstruction of the target and distractor auditory objects, formed by integrating information across frequency channels. In support of this position, Ding et al. (2014) showed that degrading the spectro-temporal fine structure (while leaving the temporal envelope intact) led to a reduction in the envelope-tracking response, suggesting that cortical envelope tracking depends on object formation. Additionally, Rimmele et al. (2015) show that degrading the spectro-temporal fine structure also leads to a reduced effect of attention on the envelope-tracking response. However, describing the effect of attention as acting on the neural representation of a formed auditory object may overlook the fact that objects, once formed need to be *maintained* over time. In other words, the object formation process must be continuously updated, and this process would be expected to require the deployment of attention to neural representations prior to their integration into a single object (Winkler et al., 2009).

For instance, before auditory objects can be formed, sounds pass through a bank of peripheral auditory filters, and the resulting spectral (frequency) channels are preserved in the ascending auditory pathway (Kaas et al., 1999; Humphries et al., 2010). These filters can introduce important non-linear transforms, including the introduction of envelopes not contained in the original stimulus (see Ghitza et al., 2013). Furthermore, attention can selectively modulate activity at even the earliest stages of encoding (Maison et al., 2001), and attention can be selectively deployed to particular frequency regions (Mondor & Bregman, 1994). For these reasons, we expect attention to be applied *non*-uniformly across spectral channels as a function of time, consistent with a model of auditory scene analysis that allows for a constant feedback loop between object formation, object selection, and low-level feature representations, with attention being able to influence the object formation process rather than just the object itself (e.g. Winkler et al., 2009).

Furthermore, we might expect that the effect of attention will be more pronounced in spectral channels corresponding to regions of the speech spectrum that are important for intelligibility, rather than simply tracking those regions that contain the most energy. Greenberg et al. (1998) suggest that 1/3-octave spectral channels in the approximately 750-2350 Hz range contribute substantially to speech intelligibility, and while this region is narrower than the speech importance region identified in the ANSI standards (ANSI, 1997), it is clear that spectral regions important for speech perception are not necessarily those that contain the most energy. This is especially true between approximately 1000 and 5000 Hz, where decreases in speech energy are not followed by decreases in speech importance.

The goal of the current study was to examine the extent to which the cortical entrainment reported in Horton et al. (2013) is selective over spectral channels. In the original study, subjects were instructed to listen to one of two speech streams presented from different free-field loudspeakers positioned 45 degrees to the left and right of center. Cortical entrainment was measured for both the attended and unattended speech stimuli as the cross-correlation between the stimulus envelope and the neural response, which recovers a temporal response function (TRF) with distinct peaks corresponding to the typical N1-P2-N2 onset response (Figure 1.1). Peaks in this temporal response function can be interpreted as delays at which the neural response reliably follows the stimulus envelope. Following the event-related potential literature (Hall III, 2007), we treat these peaks as reflecting distinct neural processes, with later peaks reflecting downstream processes in the cortical auditory hierarchy. As the auditory signal is processed downstream, information is integrated across spectral channels auditory objects are formed (e.g. Rauschecker & Tian, 2000), and since we are examining within-channel processes, we expect find larger effects at earlier latencies (i.e. N1).

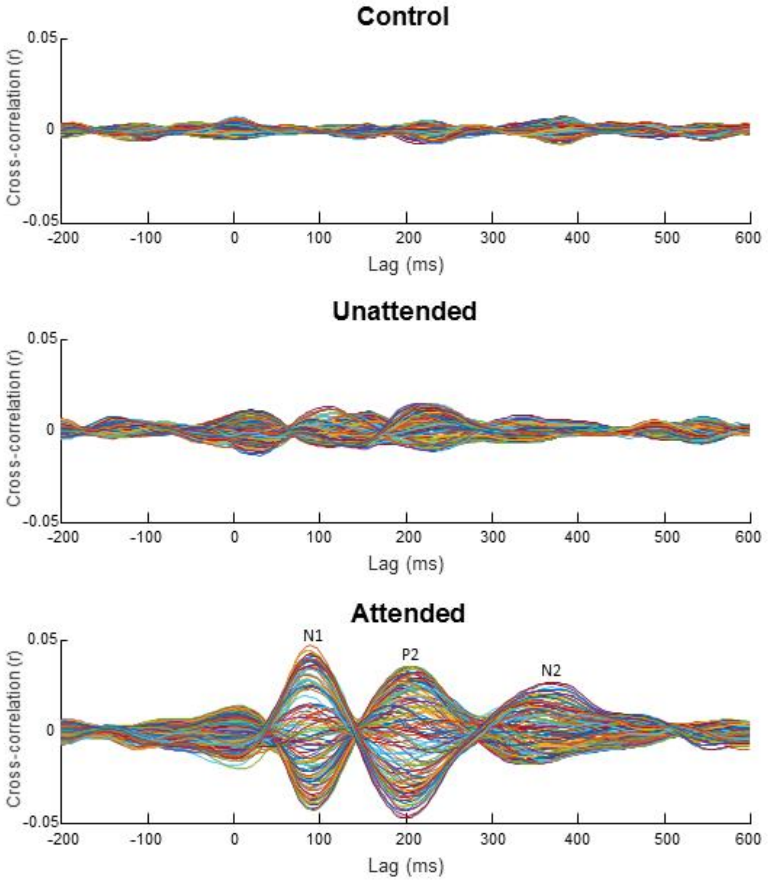


Figure 1.1: Cross-correlations between original speech envelopes and EEG activity at 128 recording channels. While we computed cross-correlations with delays from -1000 to +1000 ms, we show only -200 to +600 ms for viewing convenience, and because no significant peaks exist outside of this range. These cross-correlations generate temporal response functions that recover the N1-P2-N2 auditory evoked response, and while the response is clearest in the attended TRF, this pattern can also be observed in the unattended TRF.

In order to decompose the stimulus into spectral channels, we passed the stimulus through a gammatone filterbank with eighteen filters equally spaced on a log scale between 100 to 6246 Hz (Figure 1.2). At the output of each of these gammatone filters, which are designed to model cochlear filtering, the attended and unattended envelopes were extracted and cross-correlated with the neural response to obtain attended and unattended temporal response functions for each spectral channel.

By focusing on the *ratio* between the attended and unattended envelope-tracking response, we show that the effect of attention on the envelope-tracking response is not uniform across spectral channels in the N1 latency range. This suggests that attention is modulating the envelope-tracking response *within* spectral channels, and is therefore influencing the process of object formation rather than simply applying gain to the object itself. Furthermore, we show significant attentional modulation at high frequencies (1851 – 6246 Hz) where energy is relatively sparse, suggesting that attention is directed to high-importance rather than high-energy regions.

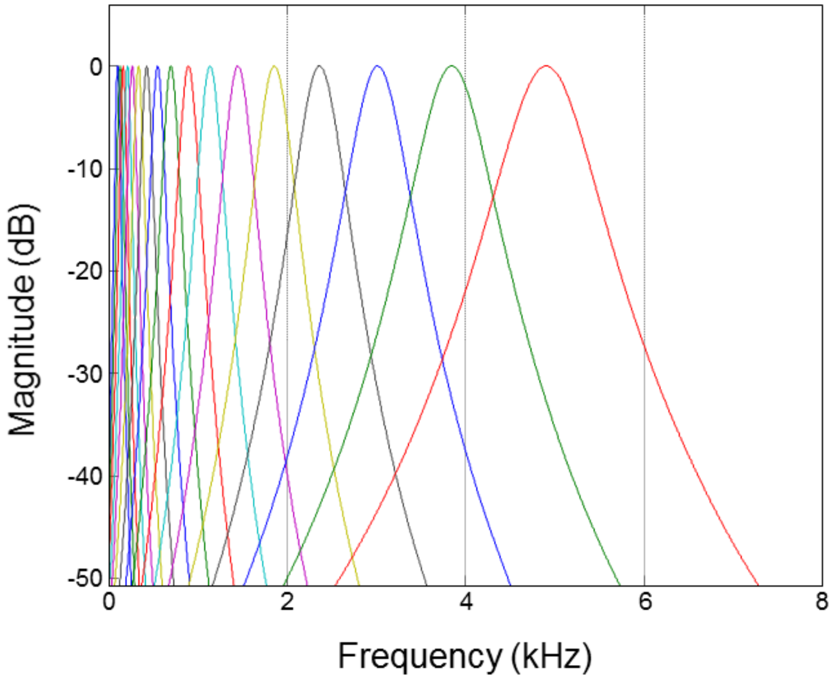


Figure 1.2: The frequency response functions of the gammatone filters used in the experiment. On a linear frequency axis, bandwidths increase with increasing center frequency. The matlab code used to generate the gammatone filter coefficients was derived from Slaney (1993).

## METHODS

The goal of the current study is to examine the extent to which the cortical entrainment reported in Horton et al. (2013) is selective in frequency. The following provides a brief description of the methods reported in Horton et al. (2013), and a detailed description of those methods novel to this study.

### *Participants*

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Ten young adults (2 female; age: 21-29) participated in the study, although one had to be excluded due to excessive EEG artifacts.

### *Task*

Each participant sat in a sound-attenuated testing chamber and faced a computer monitor that was flanked on either side by a loudspeaker. At the start of each trial, the subject was presented with a visual cue to attend to either the left or right speaker (chosen at random) while maintaining visual fixation on a cross in the center of the monitor. On each trial, two independent series of spoken sentences were played from two loudspeakers separated at a 90 degree angle. To build these speech stimuli, sentences were drawn at random from the TIMIT speech corpus (Garofolo et al., 1993) and concatenated until the total length of each speech stimulus exceeded 22 seconds. At the end of each trial, subjects were shown the transcript of a sentence from the trial, and were asked to indicate via a button press whether the sentence was played on the attended side. Subjects completed 320 trials (8 blocks, 40 trials per block), with the exception of one subject who only completed 240 trials due to equipment failure.

*EEG Recording and Pre-Processing*

High-density EEG (128 channels) was recorded with equipment from Advanced Neuro Technology. Electrodes were placed following the international 10/5 system (Oostenveld & Praamstra, 2001), and all channel impedances were kept below 10 kΩ. The EEG data was average-referenced and filtered offline with a passband of 2 to 40 Hz. The filtered data were then down-sampled from 1024 Hz to 256 Hz and segmented into individual trials which were 20 seconds long, beginning one second after the onset of the sentences. This delay was incorporated to remove any effect of a synchronous onset between the left and right speech stimuli.

*Gammatone Filtering and Envelope Extraction*

In order to simulate frequency selectivity of the auditory system, each speech stimulus was passed through a gammatone filterbank (Slaney, 1993), a well-established model of peripheral auditory filtering (for a recent review, see Lyon et al., 2010). Shown in Figure 1.2, center frequencies of the filters were equal-log-spaced from 100 to 6246 Hz (18 total filters). To extract the envelope, the output of each filter was then Hilbert transformed, and its magnitude was low-pass filtered (30 Hz). The resulting envelope was then high-pass filtered at 2 Hz to remove the DC component. Artifacts were removed using the Infomax ICA algorithm from the EEGLAB toolbox (Delorme and Makeig, 2004). The speech stimuli from the TIMIT database were originally sampled at 16000 Hz, so center frequencies close to the Nyquist rate (8000 Hz) were not considered.

*Cross-correlation analysis*

The neural response to speech stimuli was quantified by computing the cross-correlation functions between the EEG and the envelopes of the attended and unattended speech stimuli in each gammatone-filtered frequency band (see Ahissar et al., 2001 & Power et al, 2012). The cross-correlation function measures the similarity between two discrete signals $f$ and $g$ over a range of delays $n$.

$$(f \star g)(n) = \sum_{m=-\infty}^{\infty} \frac{f[t]g[n+t]}{std(f)std(g)}.$$

Since the cross-correlation is normalized between 0 and 1, the absolute magnitudes of $f$ and $g$ are not reflected in $(f \star g)$. The cross-correlation functions between the EEG and the stimulus envelope strongly resemble the N1-P2-N2 response of a typical auditory evoked potential (AEP), which is consistent with our expectation that the AEP reflects the basic response characteristics of the auditory system (Figure 1.1). For every trial, for each subject, recordings from each channel of the EEG was cross-correlated with both the attended and the unattended stimulus envelopes, and cross-correlation values were Fisher z-transformed to approximate a normal distribution, following the analysis in Horton et al. (2013).

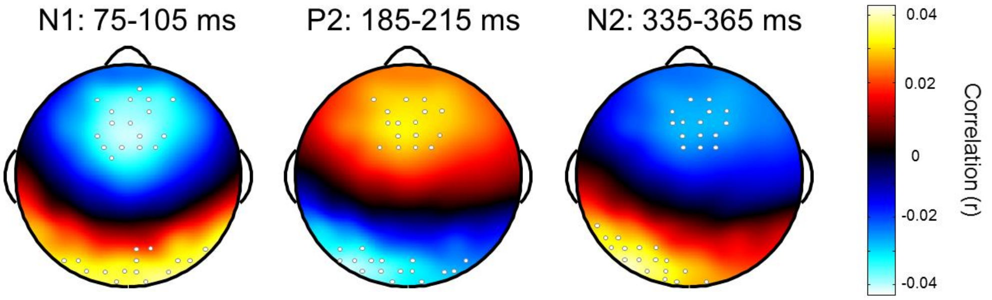


Figure 1.3: Scalp topographies for cross-correlation values at all recording sites averaged over latency ranges corresponding to N1, P2, and N2. For each range, a clear anterior-posterior dipole is observed.

Cross-correlation functions were then averaged across trials, and maximum values were extracted in each of three latency ranges that corresponded to cross-correlation peaks that resembled a typical AEP. Thirty-two out of 128 channels with the largest attended cross-correlation values were identified from grand-averaged subject data separately for three latency ranges corresponding to peaks in the AEP (labeled N1, P2, N2). A large number of channels (32) were included so that broad activity on both sides of the dipoles, shown in Figure 1.3, could be captured. Having chosen these 32 channels, for each subject, the mean of the absolute value of the attended and unattended cross-correlation functions were computed in each latency range (90 ± 25 ms, 200 ± 25 ms, and 350 ± 25 ms). Taking the absolute value allowed us to average across channels without respect to polarity. From this "composite" channel, the maximum value was selected for both listening conditions (attended, unattended) in all three latency ranges for each of the 18 gammatone-filtered envelopes and the unfiltered envelope.

To estimate a noise floor for these maxima, a bootstrap simulation was performed. A control distribution was constructed by replacing the attended and unattended stimuli on each trial with random stimuli not presented on that trial, and performing the same analysis just described over 1000 iterations. This control was useful because it shared all of the spectral and temporal characteristics of the attended and unattended envelopes but was unrelated to that particular trial's stimuli. Therefore, any nonzero values in the control cross-correlations were due purely to chance. Maximum values were considered significantly non-zero if they fell outside the 99.5th or 0.5th percentiles of this distribution.

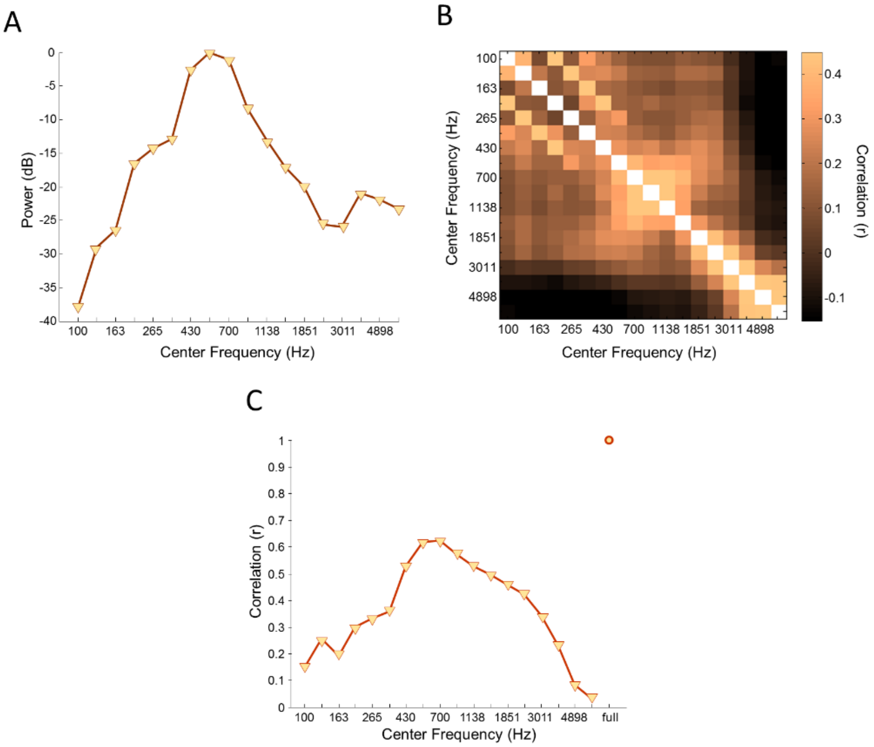


Figure 1.4: (**A**) Total power in each gammatone filter. (**B**) Correlation (normalized covariance) matrix for the same data. (**C**) Correlations between the envelope at the output of each gammatone filter with the original (full-band) stimulus envelope. This stimulus-to-stimulus correlation function can be thought of as the shape of the expected stimulus envelope-to-EEG cross-correlation by frequency function if the full-band stimulus envelope were being entrained.

### *Stimulus Power Spectrum*

Average power at the output of each gammatone filter (before envelope extraction) is shown in Figure 1.4a. We see that power increases sharply on a log scale from low frequencies to a peak at around 600 Hz (mid frequency), and decreases at high frequencies. While this pattern is roughly quadratic, we want to point out the asymmetry of the low and high-frequency tails. Specifically, the lowest frequency filters have nearly half the power as highest frequency filters.

*Stimulus-to-Stimulus Correlations*

The main goal of this analysis was to explore the extent to which envelopes extracted from peripheral channels are tracked in the cortex in a selective attention task. However, since the envelopes at the output of each filter are not uniformly correlated with the full stimulus envelope, any effect of center frequency on the strength of correlation might merely reflect the extent to which cortical tracking of the full stimulus envelope is correlated with the envelopes at different center frequencies. Shown in Figure 1.4c, correlations were highest between ~500-700 Hz, dropping off at higher and lower center frequencies. Note that, due to overlap between adjacent gammatone filters (Figure 1.2), neighboring envelopes tend to be correlated with one another (Figure 1.4b).

If the envelope-tracking response we observe is indeed a tracking of the full-band envelope, we expect that the shape of *both* the attended an unattended cross-correlation-maximum-by-frequency functions follow the stimulus-to-stimulus-correlation function in Figure 1.4c. As we have quantified the effect of attention in our analysis as the log-ratio between the attended and unattended cross-correlation maxima, we may restate this expectation as a prediction that the attended/unattended log-ratio-by-frequency function will be flat.

*Filtered Cross-Correlation Functions*

Figure 1.5 shows the cross-correlation functions for the attended, unattended, and control stimuli in four individual frequency channels (center frequencies range from 207 – 4000 Hz). As in Figure 1.1, which showed the cross-correlation functions for the attended, unattended, and control *full-band* stimuli, distinct peaks in the temporal structure can be observed in the both the attended and unattended cross-correlation functions for each frequency channel.

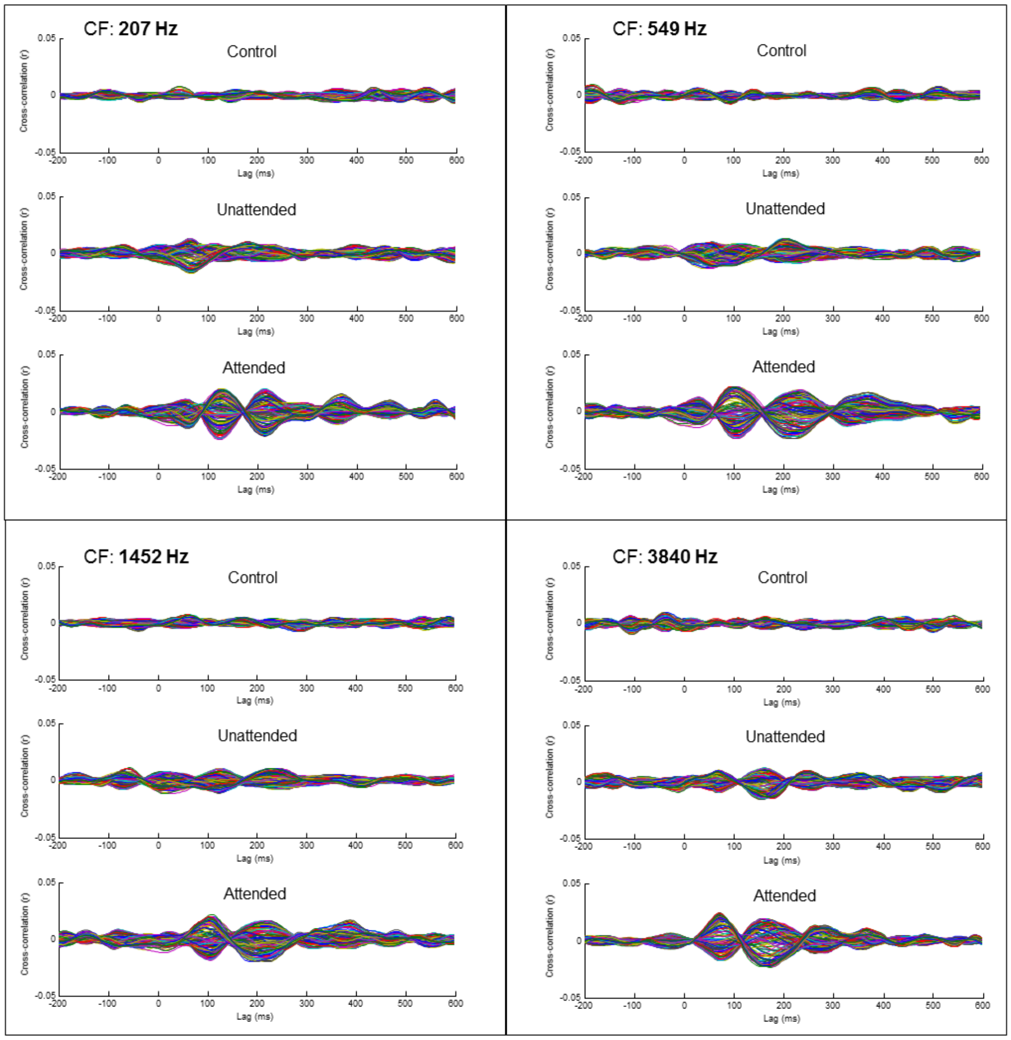


Figure 1.5: Cross-correlation functions between original speech envelopes and EEG activity at 128 recording channels for four representative frequency channels (CF) that span the range of CFs included in our analysis. Notice that both the attended and unattended cross-correlation functions show significant structure in the ~65-365 ms latency range, while the control cross-correlation functions do not.

*Statistical Procedure*

Our choice of statistical procedure was motivated by two concerns. First, our independent measures had a covariance structure that was not compound symmetric (Figure 1.4b). Second, the number of independent measures (i.e. gammatone filters, see Figure 1.2) was selected somewhat arbitrarily, albeit with the goal of maximizing coverage of the spectrum while minimizing filter overlap. In other words, our decomposing of the speech stimulus into eighteen different spectral bands was simply a convenience, and runs the risk of artificially inflating the number of independent measures we use in our analysis.

We first considered using a linear mixed-effects model that allowed us to specify an autoregressive covariance structure for the fixed effect of center frequency. However, this model, like the ANOVA, adjusts degrees of freedom based on number of independent measures, so even if we accounted for the covariance structure, such an approach runs the risk of artificially inflating degrees of freedom and thus artificially inflating significance.

We decided it was more reasonable to reduce the data into three independent spectral channels instead of eighteen, as this would alleviate both concerns. First, by collapsing (averaging) over low, mid, and high frequencies (on a log scale), we restrict the covariance problem to two borders of these three frequency regions. Second, we reduce the number of independent measures down to three, which is far more conservative in terms of degrees of freedom, and allows us to run a standard multivariate ANOVA (Vasey & Thayer, 1987).

Furthermore, a collapse over low, mid, and high frequencies follows naturally from three aspects of the natural speech. First, the power spectrum of the speech (e.g. Figure 1.4a) has relatively less energy at low and high frequencies, with most of the energy in the mid frequencies, and if we expect the envelope-tracking response to follow stimulus energy, such a division is appealing. Second, fundamental frequencies for adult male and female talkers do not typically exceed 300 Hz, and first formants do not typically fall below 400 Hz, establishing a natural point of division between low and mid frequencies (Titze, 1994). Third, while there is substantial overlap between frequency regions important for the perception of vowels and consonants, spectral information in the 400-1500 Hz range is crucial for the perception of vowels, and bursts of frication in the 1500 Hz and above range are crucial for consonant identification (see Li, Menon & Allen, 2010), again forming a somewhat natural division between mid and high frequencies. Therefore, we collapsed across the lower six (100 – 338 Hz), the middle six (430 – 1452 Hz), and the highest gammatone filters (1851 – 6246 Hz), effectively reducing the data from eighteen independent frequency regions down to three (low, mid and high).

Using these frequency ranges (spectral channels), we ran a 2-factor (low/mid/high x attended/unattended) multivariate ANOVA. We used a multivariate approach because this allows us fit the covariance structure empirically, rather than assuming compound symmetry (Vasey & Thayer, 1987). This analysis was run separately for the N1, P2 and N2 latency ranges. In ranges that revealed a significant interaction between frequency and attention, post-hoc analyses were performed on the attended/unattended log-ratio. We decided to characterize the effect of attention as an attended/unattended ratio because it effectively removes any effect of the envelope-tracking response not due to attention. In other words, both attentional enhancement of the target and attentional suppression of the masker will be reflected in the ratio, and it is the relative strength of the attended and unattended envelope-tracking response in each frequency region that best summarizes the effect of attention. The log-transform was applied so that the distribution of ratios was approximately normal.

## RESULTS

Latency ranges for the N1, P2, and N2 peaks were defined, and a subset of maximally-responding channels were selected to form a region of interest (ROI) within each latency range (Figure 1.3). Attended and unattended cross-correlation maxima were then selected from the ROI time series, yielding the functions shown in Figure 1.6a. A bootstrap was performed to estimate a noise floor for cross-correlation maxima due to chance (see Experimental Procedures). For the purposes of statistical analysis, we collapsed over low, mid, and high frequency regions, shown in Figure 1.6b. In latency ranges where a significant interaction between attention (attended vs unattended) and frequency region was observed, we quantified the effect of attention as the attended/unattended log-ratio, and performed post-hoc tests on this log-ratio function (Figure 1.7). We focus on the ratio because this provides a summary effect of attention, reflecting both target (attended stimulus) enhancement and masker (unattended stimulus) suppression.

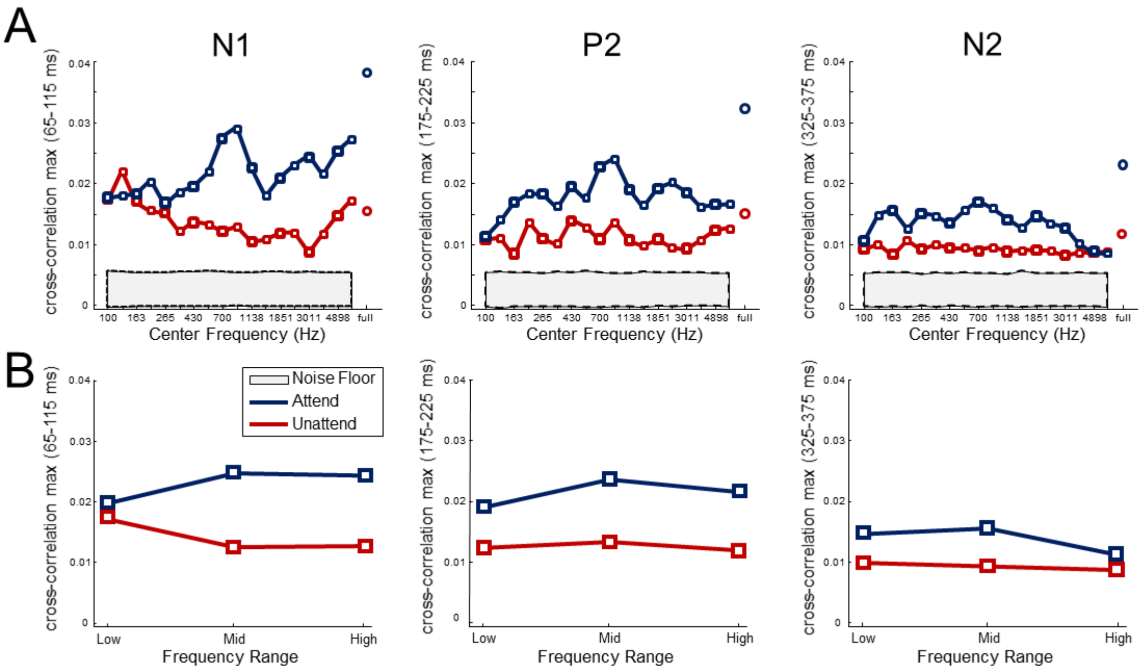


Figure 1.6: (**A**) Cross-correlation maxima as a function of gammatone filter center frequency for latencies corresponding to the N1 (90 ± 25 ms), P2 (200 ± 25 ms) and N2 (350 ± 25 ms) peaks. The noise floor (gray) shows the range of correlation values that would occur by chance if the stimulus envelope is unrelated to the EEG. (**B**) The data-reduced version of (A), collapsed into low (100 – 338 Hz), mid (430 – 1452 Hz), and high (1851 – 6246 Hz) frequency regions.

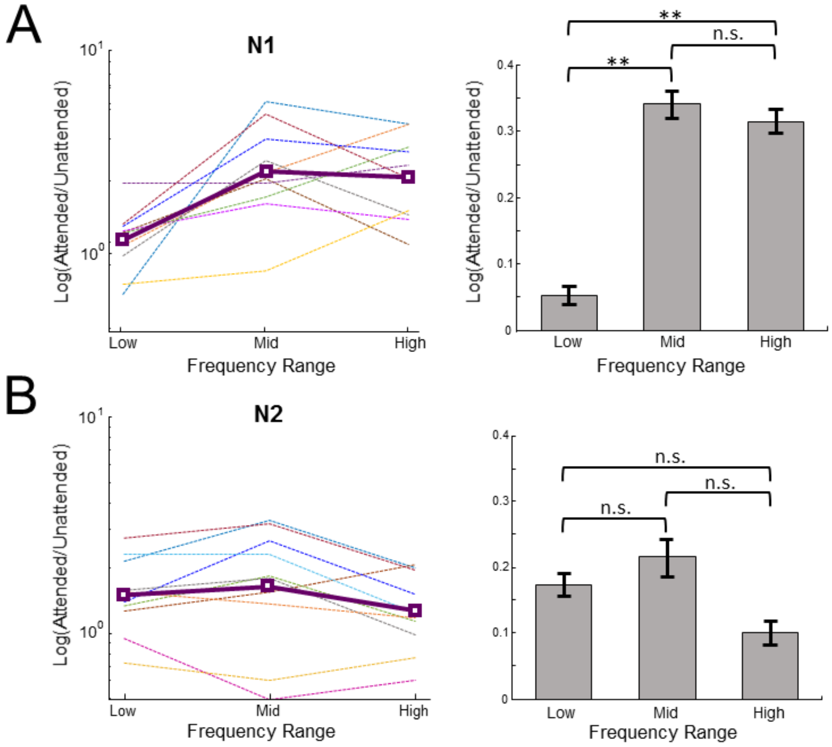


Figure 1.7: (**A**) Log-ratios between attended and unattended cross-correlation maxima as a function of frequency region (Low: 100–338 Hz; Mid: 430–1452 Hz; High: 1851–6246 Hz) in the N1 latency range. The solid line indicates the grand average, and each individual dotted line represents an individual subject. On the right of this plot is a bar graph showing the outcome of paired-comparison post-hoc tests. Error bars represent standard errors of them mean. (**B**) Same as (**A**) but for the N2 latency range.

*The N1 latency range*

A 2-factor MANOVA in the N1 latency range revealed a significant interaction between frequency and attention (Pillai's trace = .709, $F[1,9] = 9.7$, $p = .007$, $\eta_p{}^2 = .709$). Therefore, we examined the simple effect of frequency for both the attended and unattended functions. This post-hoc MANOVA analysis revealed a marginally significant (after Bonferroni correction) simple effect of frequency for the attended function (Pillai's trace = .534, $F[2,8] = 4.59$, $p = .047$, $\eta_p^2 = .534$), and a significant simple effect for the unattended function (Pillai's trace = .815, $F[2,8] = 17.6$, $p = .001$, $\eta_p^2 = .815$).

Having found a significant interaction, we performed paired-comparisons on the attended/unattended log-ratio, which revealed that the envelope-tracking response is significantly smaller in the low frequency region than in the mid ($p = .004$) and high frequency regions, ($p = .005$), but responses are not significantly different between mid and high frequency regions ($p = .71$). These results are shown in Figure 1.7a.

*The P2 latency range*

A 2-factor repeated measures MANOVA in the P2 latency range revealed a significant main effect of attention (Pillai's trace = .702, $F[1,9] = 21.2$ $p < .001$, $\eta_p^2 = .702$), a significant main effect of frequency region (Pillai's trace = .611, $F[2,8] = 6.29$, $p = .023$, $\eta_p^2 = .611$), but no significant interaction between frequency region and attention (Pillai's trace = .211, $F[2,8] = 1.07$, $p = .387$, $\eta_p^2 = .211$). Due to a lack of a significant interaction between frequency and attention, we did not perform post-hoc analyses on the attended/unattended log-ratio. However, we further investigated the significant main effect of frequency region by examining the simple effect of frequency for both the attended and unattended functions. We found that neither the attended (Pillai's trace = .455, $F[2,8] = 3.34$, $p = .088$, $\eta_p^2 = .455$) nor unattended (Pillai's trace = .110, $F[2,8] = .49$, $p = .627$, $\eta_p^2 = .11$) function reached significance.

*The N2 latency range*

A 2-factor repeated measures MANOVA in the N2 latency range revealed a significant main effect of attention (Pillai's trace = .495, $F[1,9] = 8.81$ $p = .016$, $\eta_p^2 = .495$), no significant main effect of frequency region (Pillai's trace = .434, $F[2,8] = 3.06$, $p = .103$,

$\eta_p{}^2$ = .434), and a significant interaction between frequency region and attention (Pillai's trace = .668, $F$[2,8] = 8.05, $p$ = .013, $\eta_p{}^2$ = .668). Therefore, we performed paired-comparisons on the attended/unattended log-ratio but found no significant comparisons (all $p$ > .05). These results are shown in Figure 1.7b. Furthermore, effect sizes were moderate for the low-frequency to high-frequency ($d$ = .422) and mid-frequency to high-frequency ($d$ = .491) comparisons, suggesting that small effect sizes are not driving the lack of significance.

## DISCUSSION

The data reported here suggest that the modulatory effects of attention on the neural tracking of speech envelopes can depend on frequency region. The summary effect of attentional modulation was quantified as the ratio between attended and unattended cross-correlation maxima. This ratio showed a significant effect of frequency region in the N1 latency range, but not in the P2 and N2 latency ranges.

### *The effect of attention in the N1 latency range*

Significant differences in the attended/unattended ratio across frequency regions in the N1 latency range are inconsistent with a model of envelope-tracking that strictly follows the full-band envelope, as such a model predicts that the attended/unattended ratio across spectral channels would remain constant (or flat). This conclusion is further supported by the fact that the simple main effect of frequency region was significant for the unattended function and marginally significant (after Bonferroni correction) for the attended function.

Specifically, mid and high frequency regions show significantly greater attentional modulation than low frequency regions (Figure 1.7a). If we consider that the stimulus power spectrum (Figure 1.4a) peaks at mid frequencies, there is an intuitive interpretation of the difference between low and mid frequencies, namely, that attention is deployed in mid frequency channels because these channels contain the most stimulus energy.

If attentional modulation of the envelope-tracking response were simply following stimulus energy however, we would expect to see a difference in the attended/unattended ratio between mid and high frequency regions. The fact that the attended/unattended ratio in the high-frequency region is significantly larger than in the low-frequency region and *not* significantly different than the mid-frequency region suggests that attentional modulation is not following stimulus energy in the high frequency region. If we consider that fricatives provide high-frequency, broadband bursts of energy (Strevens, 1960), and that the cortex is prone to respond to abrupt onsets (Phillips, Hall & Boehnke, 2002), it is perhaps not surprising that attention would be directed to those channels that carry these abrupt onsets, namely, those in the high-frequency region. Recent studies have demonstrated that the timing and frequency content of fricative bursts at an above ~1500 Hz are crucial for differentiating phonemes (Li, Menon & Allen, 2010; Li et al., 2012). We might also think of these fricative bursts as acoustic landmarks for syllable structure and word boundaries, and therefore particular important in degraded listening environments (Li & Loizou, 2008). Indeed, Doelling et al. (2014) showed that sharp envelope fluctuations drive envelope tracking, and that this tracking correlates with intelligibility.

*Effects of attention in the P2 and N2 latency range*

In the P2 latency range, our analysis did not reveal a significant interaction between frequency region and attention, which is to say that the effect of frequency region was not significantly different between the attended and unattended envelope-tracking response. However, in the N2 latency range, a significant interaction was observed, which prompted us to analyze the attended/unattended ratio. Shown in Figure 1.7b, no pairwise comparisons were significant, which limits our ability to discuss this interaction. While it may be the case that the attended/unattended ratio depends on frequency region, this effect was not robust in our dataset.

Instead, our data suggest that in the N1 latency range, there is a robust difference between the effect of frequency region on the attended and unattended envelop-tracking response, while in the P2 and N2 latency range this difference was not observed. Later latency ranges (P2, N2) reflect downstream processes in the cortical auditory hierarchy (Shahin et al., 2005; Tonnguist-Uhlen, 1996). As the auditory signal is processed downstream, information is integrated across spectral channels (i.e., auditory objects are formed; Rauschecker & Tian, 2000). Therefore, the pattern of attention effects observed here (frequency-specific in N1, non-specific in P2, N2) may reflect a transition from a low-level, tonotopic representation of the signal (N1) to a high-level, object-based representation of the signal (P2, N2).

*Attentional enhancement vs. suppression*

We have chosen to focus our discussion thus far on the attended/unattended ratio, as this is a summary effect of attention that can reflect both target (attended stimulus) enhancement and masker (unattended stimulus) suppression. This is motivated in part from a lack of control shape against which to test our attended and unattended envelope-tracking responses across frequency region. However, as shown in Figure 1.6b, we see that while the attended function rises from low to mid/high frequencies (in the N1 latency range), the unattended function falls. We believe such an effect is consistent with suppression of the competing talker, especially if we consider the attended function as a proxy for a control (Horton et al., 2013). In particular, we might expect greater envelope tracking in the mid-frequency region relative to the low-frequency region, as there is far more energy in the mid-frequency region. If such an assumption is valid, then the fact that the envelope tracking response to the competing talker *decreases* from the low-frequency to mid-frequency region almost certainly reflects attentional suppression.

*Contrast to previous research*

There are two results that should be considered relative to the findings reported in the current study. First, Ding & Simon (2012a) failed to find an effect of attention on the shape of the spectral response function, which plots correlation as a function of frequency. In other words, the ratio between the attended and unattended envelope-to-MEG correlations was flat. However, there are a number of differences between our study and theirs. Perhaps most importantly, their data were reported after projecting the data onto a single source, which implicitly filtered the MEG time series. We made no attempt to localize a single source, and our data almost certainly include activity from multiple sources within and outside of auditory cortex (Giard et al., 1994). Furthermore, our cross-correlation analysis independently computed a temporal response function for each frequency channel, while the spectro-temporal receptive field (STRF) analysis used by Ding & Simon (2012a) fit temporal and spectral response functions with the same model. Second, Mesgarani & Chang (2012), recording ECoG from electrodes on the surface of the superior temporal gyrus (the location of A2), found that the effect of attention was spatially distributed among recording sites, and did not identify any particular regions that were driving the attentional modulation. This means that, to the extent that activity in A2 is tonotopically organized, the effect of attention is distributed rather than localized in frequency. The distribution of the effect of attention however, was not statistically evaluated, and it is therefore difficult to make direct comparisons to our result.

With these results in mind, it is possible that attentional modulation of the envelope-tracking response can occur both within and across spectral channels. Indeed, there is no reason to assume that envelope tracking within spectral channels *precludes* envelope tracking to the full-band (or integrated) envelope.

However, it should also be noted that Lakatos et al. (2013) reported entrainment effects within specific tonotopic regions of A1, and reported that frequency regions containing the target entrained to the stimulus with high-excitability phases, while frequency regions distant from the target entrained with low-excitability phases. This suggests, at the very least, that entrainment can reflect attentional dynamics within frequency channels, and seem to suggest a mechanism by which entrainment can selectively suppress the output of unattended/ignored channels.

### *Limitations and Suggestions for Future Research*

We report that the effect of attention on the neural response to the speech envelope can be frequency dependent, though our analysis only permits a narrow interpretation of this dependency. Because we did not systematically vary the frequency content of our speech stimuli over trials, we cannot suggest that attention is allocated to different frequency bands on a trial-by-trial (or utterance-by-utterance) basis. Instead, our results only suggest that on average, in the N1 latency range, attention modulates the neural envelope-tracking response in a frequency-dependent fashion (Figure 1.7a). This frequency dependency may reflect a fixed property of the auditory system, or it may represent the average response of utterance-specific attentional modulation. In other words, we don't know whether or not the frequency dependency we observe represents an active tracking of the frequency content of each utterance. Furthermore, our analysis does not rule out the possibility that attention is modulating the neural response to the full-band envelope of the integrated auditory object in addition to modulating the envelope-tracking response within individual frequency channels. Indeed, as explained above, there is no reason to suspect that attentional modulation of the envelope-tracking response may occur *within* and *across* spectral channels. The first of these limitations can be addressed with a follow up study that systematically fixes the spectra of the attended and unattended speech stimuli across trials, and while the second limitation may prove difficult to address with EEG, techniques with better spatial resolution may be able to resolve this issue.

### *Conclusions*

In a multi-talker listening environment, the envelope-tracking response to the attended talker is larger than the response to the unattended talker in three latency ranges corresponding to the N1, P2, and N2 peaks in the auditory evoked response. Crucially, in the N1 latency range, attention differentially modulates the envelope-tracking response in different frequency regions, suggesting that attention is deployed differentially across spectral channels. This result is inconsistent with the suggestion that attention is deployed exclusively to the envelope of an integrated auditory object, and instead suggests that attention influences the process of object formation.

# CHAPTER 2

## INTRODUCTION

Synchronous oscillations in the cortex arise naturally from both synaptic delays in local neural circuits and transmission delays in long-range neural connections (Nunez & Srinivasan, 2006). For certain neural circuits, it has been demonstrated that oscillatory activity reflects synchronized modulations of neuronal excitability (Buzsaki & Draguhn, 2004). Based in large part on these findings, and on findings that attention can lead to increased neural synchrony (e.g. Fries et al., 2001), it has been suggested that oscillations enable coordination at different timescales across and within cortical networks, and in this way govern perceptual and cognitive processes (e.g. Fries, 2005). Lending support to this claim, recent studies have shown that cortical oscillations can hierarchically couple such that the phase of lower-frequency oscillations can modulate the power at higher-frequency oscillations, and that this coupling is task-specific (Lakatos et al., 2005; Canolty et al., 2006; Lakatos et al., 2008). While this literature has generated some claims that remain controversial, it has also proposed a set of neural-mechanisms through which oscillations can govern perceptual processes.

It has been proposed that phase-resetting cortical oscillations support speech perception by aligning fluctuations of neuronal excitability to periodic changes in the speech stimulus, a process that has been labelled cortical entrainment (Ding & Simon, 2014). Specifically, cortical entrainment is thought to support the parsing of acoustic input into discrete linguistic units such as phrases, syllables, and phonemes by aligning hierarchically-coupled oscillations at multiple time-scales to the natural fluctuations of the speech stimulus (e.g. Ghitza, 2011; Giraud & Poeppel, 2012). Supporting this claim, Ding et al. (2016) identify a neural response that entrains to periodic fluctuations in high-level linguistic units (including verb phrases and sentences) in the absence of acoustic cues, distinct from a neural response that entrains to the acoustic envelope regardless of linguistic content. However, while this study elegantly differentiates between entrainment to linguistic and acoustic cues, the speech stimuli used to make this distinction were artificially periodic. Specifically, they played single words at a rate of 4 Hz, with phrases constituting 2-word units (2 Hz repetition rate), and sentences constituting 4-word units (1 Hz repetition rate). Modulation rates in natural speech however, are far less regular and less predictable, as are fluctuations in linguistic units such as phrases and sentences, making it difficult to generalize these results to natural speech.

Indeed, for natural speech, it has proven difficult to demonstrate entrainment to linguistic features. A number of studies have failed to show an effect of intelligibility on entrainment, suggesting that entrainment may not reflect the tracking of linguistic features. For instance, Howard & Poeppel (2010) show that entrainment to natural speech is no greater than entrainment to time-reversed speech (see also Zoefel & VanRullen, 2016; though see Hertrich et al., 2013). Furthermore, Doelling et al. (2014) showed that entrainment to click trains, which carry no linguistic information but contain maximally abrupt acoustic changes, was significantly higher than entrainment to natural speech, suggesting that entrainment reflects the abruptness of acoustic transitions rather than linguistic content.

Other studies have shown that degrading the intelligibility of speech stimulus leads to a decrease in the strength of entrainment to that stimulus. However, we know that entrainment to acoustic changes can occur in the absence of linguistic information (e.g.

Doelling et al., 2014), so when manipulations in intelligibility also fundamentally alter the acoustics of the speech stimulus, it is difficult to conclude that the effects of this manipulation on entrainment arise from the degradation of linguistic cues. For instance, Peelle et al. (2013) show that entrainment to intelligible vocoded speech (16-channels) is significantly greater than entrainment to unintelligible vocoded speech (1-channel). In this case, it is unclear whether entrainment is driven by intelligibility (and therefore reflects a linguistic contribution) because vocoding severely degrades the acoustic input to the auditory system, and this degradation is more severe for speech vocoded with lower numbers of channels (Shannon et al., 1995). It is therefore possible the decrease in entrainment response from 16-channel to 1-channel speech reflects a degradation in acoustic input rather than a degradation of intelligibility. Such a concern can also be raised when interpreting the results of Ahissar et al. (2001), who showed that entrainment to time-compressed speech was reduced relative to natural speech. Again, time-compression is a manipulation that severely distorts the acoustic input to the auditory system, making it difficult to determine whether the reduction in entrainment reflects degraded linguistic information or degraded acoustic information.

In an attempt to avoid these acoustic confounds, Millman et al. (2015) used the same two vocoded sentences before and after perceptual learning to measure the effect of intelligibility on the entrainment response. They used a 3-channel vocoder to ensure that the sentences were unintelligible before training. Prior to training, these vocoded sentences were presented over multiple trials and the MEG response was recorded. During training, one of these vocoded sentences was then paired with the non-vocoded original so that the meaning of the sentence could be understood despite the vocoding. After training, these vocoded stimuli were once again presented to the listener while the MEG response was recorded such that one vocoded sentence was now intelligible while the other remained unintelligible. They found no difference in the strength of entrainment to the post-training intelligible sentence and the entrainment response to the post-training unintelligible sentence, suggesting that entrainment is driven by speech acoustic rather than linguistic cues.

We considered it possible however, that the failure of Millman et al. (2015) to find an effect of intelligibility on the entrainment response may be due to overlearning through a use of limited speech materials. We know that the neural response to verbal materials can decrease after overlearning (Thompson & Thompson, 1964), and we thought it possible that mechanisms involved with speech comprehension may be less active after overlearning. We also considered that a lack of a demanding behavioral task may have limited the amount of attention listeners deployed in encoding the stimuli, which may also have contributed to the lack of effect. In the current study, we attempted to remove these potential confounds, while also employing a design that allowed us to contrast effects of prior knowledge and effects of intelligibility on the entrainment response.

Following Millman et al. (2015), we used tone-vocoding to degrade the intelligibility of spoken sentences. Sentences were drawn from the TIMIT database (Garafolo et al., 1993), which contains sentences from 630 different speakers of American English. Vocoding has been widely used as a procedure for manipulating intelligibility, and previous research has shown that while 16-channel vocoded speech is perfectly intelligible, 3-channel vocoded speech is largely unintelligible (Loizou et al., 1999). In order to differentially study the effect of prior knowledge and intelligibility, both 16-channel and 3-channel vocoded sentences were used.

In order to manipulate the intelligibility of 3-channel vocoded sentences without altering their acoustic properties, we primed the degraded (vocoded) sentences with non-degraded (natural-speech) sentences. Though text priming has been used in the literature (e.g. Sohoglu et al., 2014), we chose not to use text priming due to the neuroanatomical differences between reading and spoken language comprehension, particularly at early stages of acoustic/phonetic processing (Cohen et al., 2002; Buchweitz et al., 2009; Price, 2012), and the fact that we did not want cortical networks not typically involved with speech perception to influence our neural recordings. When a vocoded sentence is preceded by a non-vocoded copy of itself (valid condition), the intelligibility of the vocoded sentence is largely restored. Conversely, when the vocoded sentence is preceded by a non-vocoded sentence that is unrelated to the vocoded sentence (invalid condition), the intelligibility of the vocoded sentence is unaffected (Remez et al., 1980). We tested both of these conditions (valid and invalid).

Manipulating intelligibility using valid and invalid primes does not alter the acoustics of the target vocoded sentences. It does however, alter the a priori acoustic expectations brought to bear upon the vocoded sentences. To capture the effects of this prior knowledge, we use 16-channel vocoded speech under the same priming manipulation. Since 16-channel vocoded speech is intelligible, any difference between valid and invalid conditions can be attributed to prior knowledge. This 2 (valid vs. invalid) x 2 (3-channel vs. 16-channel) design allows us to isolate the effects of intelligibility and prior knowledge in the absence of acoustic confounds, allowing us to measure the degree to which the entrainment response to ongoing speech tracks acoustic or linguistic features of the stimulus. If the entrainment response depends on intelligibility, we expect a significant interaction between cue validity and number of vocoded channels such that a valid cue leads to a greater increase in the entrainment response to 3-channel vocoded speech than to 16-channel vocoded speech.

Finally, to motivate listeners to attend to the vocoded stimuli, after each vocoded sentence, a short clip of vocoded material (probe) was presented and listeners were asked to indicate whether or not that clip came from the vocoded sentence just heard on that trial. In an attempt to control for variation in task difficulty across conditions, an adaptive tracking procedure was used with separate tracks for each condition. Following a 2-down/1-up tracking procedure, the duration of the probe was adaptively varied according to the listener's response, converging on the duration necessary to achieve 71% correct (Levitt, 1971).

## METHODS

### *Participants*

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Fourteen young adults (6 female; age: 23-28) participated in the study, although one (male) was excluded due to excessive EEG artifacts. Participants were not screened for handedness.

### *Stimuli*

Speech materials were drawn from the TIMIT database (Garafolo et al., 1993), which contains sentences from 630 different speakers of American English. Specifically, we selected all sentences between 3.5 and 5 seconds in duration, yielding a total of 1247 sentences. Many of the TIMIT sentences contain brief silent periods at the start of the recording, and these silent periods were removed prior to testing for all sentences used in the experiment.

We used a tone-vocoder algorithm with logarithmic band-spacing developed by Nie et al. (2005) to manipulate the intelligibility of these sentences. Vocoding is a four-step process that (1) band-pass filters the input stimulus into individual frequency bands, (2) extracts a low-pass filtered envelope from the output of each filter, (3) modulates a tone carrier with the resulting envelope, and (4) band-pass filters the resulting bands and sums them together. Both 3-channel and 16-channel vocoders were used in the experiment. In our implementation of this algorithm, we specified a 30-Hz low-pass filter cutoff.

### *Task*

Each participant sat in a single-walled sound-attenuated booth and stimuli were presented diotically at 70 dB SPL over Stax electrostatic headphones. Testing was conducted over two sessions, each of which contained four blocks. For each session, two blocks contained exclusively 16-channel vocoded trials and two blocks contained exclusively 3-channel vocoded trials (order was randomized). Prior to each recording block, participants were asked to rate the perceived intelligibility of target vocoded sentences in both the valid and invalid prime conditions on a 1-6 scale, where a 6 indicated that they understood all of the words in the sentence, and 1 indicated that they understood none of the words. On valid trials (when the target and prime matched), participants were asked not to consider the intelligibility of the prime when making their intelligibility judgements. Since 16-channel vocoded speech is intelligible regardless of priming, in the interest of time, we only obtained a single intelligibility rating for each condition (valid vs.

invalid prime) prior to 16-channel experimental blocks. Prior to 3-channel blocks however, we obtained intelligibility ratings for fifteen target sentences in each condition to provide a stable estimate. Intelligibility ratings were gathered to ensure that (1) in the 3-channel vocoding condition, valid primes led to higher intelligibility ratings than invalid primes, and (2) this effect was stable over blocks. Due to the high variability of individual talkers in our stimulus set, we expected minimal adaptation to the vocoded stimuli.

On each trial, target vocoded sentences were preceded by a natural-speech prime sentence, and followed by a probe clip of vocoded speech material (Figure 2.1). On half the trials, these natural-speech prime sentences matched the target vocoded sentence (valid prime), and on the other half, they didn't match (invalid prime). Participants were asked to indicate whether or not the probe was drawn from the target vocoded sentence, and the duration of the probe was adaptively varied following a 2-down/1-up procedure, which converges on the duration necessary to achieve 71% correct (Levitt, 1971). We used a ratio step-size of 1.2 such that increases in probe length were in steps of 1.2 times the previous length, and decreases were in steps of 1/1.2 (or 0.83) times the previous length. Within each block of 100 trials, separate adaptive tracks were used for trials with valid and invalid primes, of which there were 50 for each trial type. Across all trials and all blocks, prime and target sentences were drawn without replacement from the experimental subset of the TIMIT database, and sentences were drawn at random for each participant. EEG was recorded simultaneously with this task. While we used an adaptive track in attempt to avoid confounds of unequal effort across conditions, we did not employ an objective measure of attention.

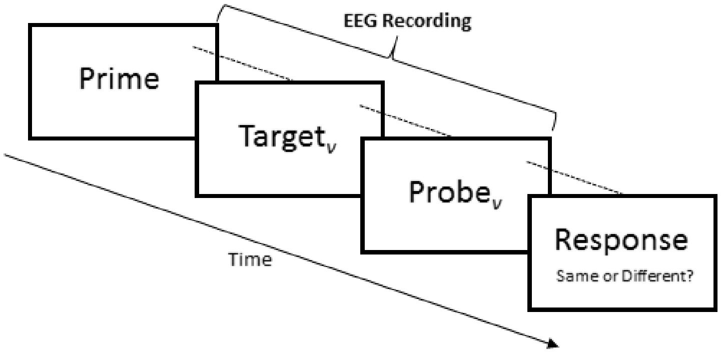


Figure 2.1: Schematic of an experimental trial. A vocoded target is *preceded* by a natural-speech (non-vocoded) cue, and *followed* by a vocoded probe. Participants are asked to indicate whether (same) or not (different) the probe snippet was drawn from the target. On half the trials the cue and the target were *matched (valid)*, and on the other half they were *mismatched (invalid)*.

*EEG Recording and Pre-Processing*

High-density EEG (128 channels) was recorded with equipment from Neuroscan. Electrodes were placed following the international 10/5 system (Oostenveld & Praamstra, 2001), and all channel impedances were kept below 10 kΩ. The EEG data was sampled at 1000 Hz, and filtered offline with a passband of 1 to 50 Hz. The filtered data were then segmented into individual trials which were 3 seconds long, beginning 500 ms after the start of the sentence. This delay was incorporated to remove the onset response to the start of the sentence. Artifacts were removed from the segmented EEG data using the Infomax ICA algorithm from the EEGLAB toolbox (Delorme and Makeig, 2004).

*Envelope Extraction and Cross-Correlation Analysis*

The envelopes of the speech materials were extracted, band-pass filtered from 1 to 50 Hz, and down-sampled to 1000 Hz. The first 500 ms of this envelope was removed and the subsequent 3 s retained to align with the EEG data. This procedure was performed for both the natural-speech primes and the vocoded targets. Paired with the neural response, this broadband speech envelope submitted to a cross-correlation in order to quantify the entrainment response to speech. This cross-correlation was performed for both the vocoded target sentences and the natural-speech prime sentences for all recording channels (Figure 2.2). The cross-correlation functions between the EEG response and the stimulus envelope peak at an average latency of 75 ms, slightly earlier than the typical N1 response of an auditory evoked potential. For ease of discussion however, we will refer to this peak as an N1, since it occurs within the N1 range and since the cross-correlation function is likely to reflect a contribution from typical N1 generators. For every trial, recordings from each channel of the EEG were cross-correlated with both the target and the prime sentence envelopes, and cross-correlation values were Fisher z-transformed to provide an approximately normal distribution, following the analysis in Baltzell et al. (2016).

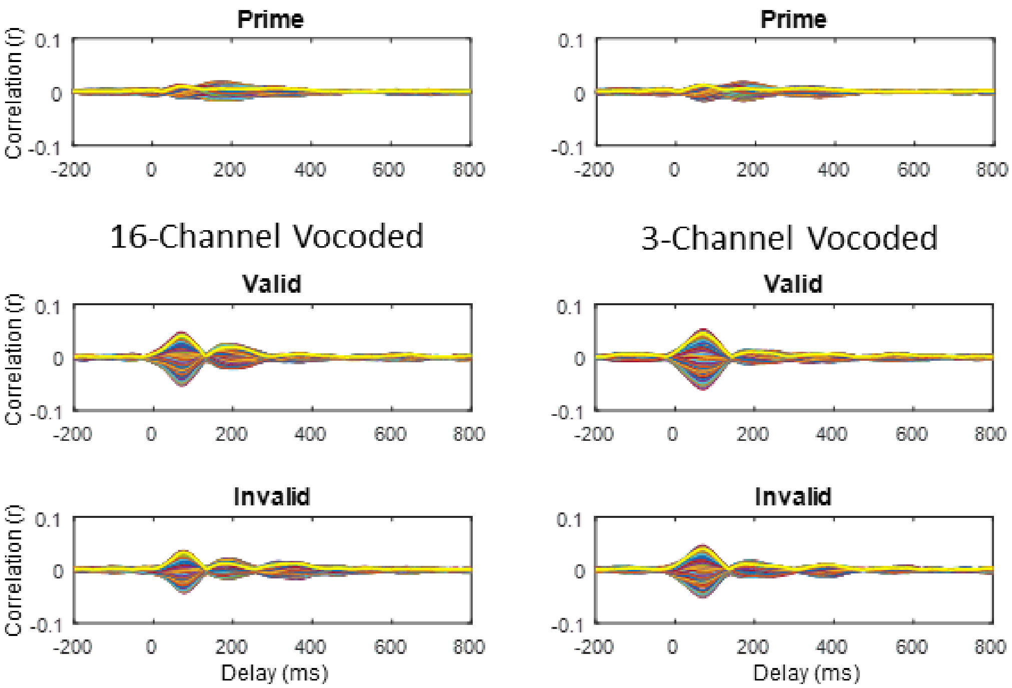


Figure 2.2: Grand-averaged cross-correlation functions (bandpass filtered envelopes [1-50 Hz] were cross-correlated with the EEG recorded from that trial) for the natural-speech cues and the vocoded targets. Thick yellow lines indicate the time course of the ROI. The only prominent peak in these functions is contained within a latency range of ~25 to ~125 ms.

Cross-correlation functions were then averaged across trials and subjects to form a grand average, and mean values were extracted in the N1 latency range for each channel, which extended from 25 to 125 ms. The absolute value of these means was computed so that a region of interest (ROI) could be defined without respect to polarity. We chose to analyze mean rather than peak cross-correlation values so as to correct for any noise in the estimation of the neural delay. This noise may be physiological, or it may be an artifact of the high variability of our speech stimuli. To define the ROI we selected the thirty electrode channels with the highest mean cross-correlation in the N1 latency range for all sentence types. These six sentence types included the natural-speech prime and the vocoded targets in the valid and invalid prime conditions (prime/valid/invalid) for both the 3-channel and

16-channel vocoder conditions. Of this set of 180 (six times thirty) channels, duplicates were removed leaving only 44 unique channels, and these 44 channels comprised our ROI (Figure 2.3). Having defined our ROI from the grand-averaged data, for each subject, the mean of the absolute value of the cross-correlation functions for each sentence type were computed first over the ROI, and then over the N1 latency range (25 to 125 ms).

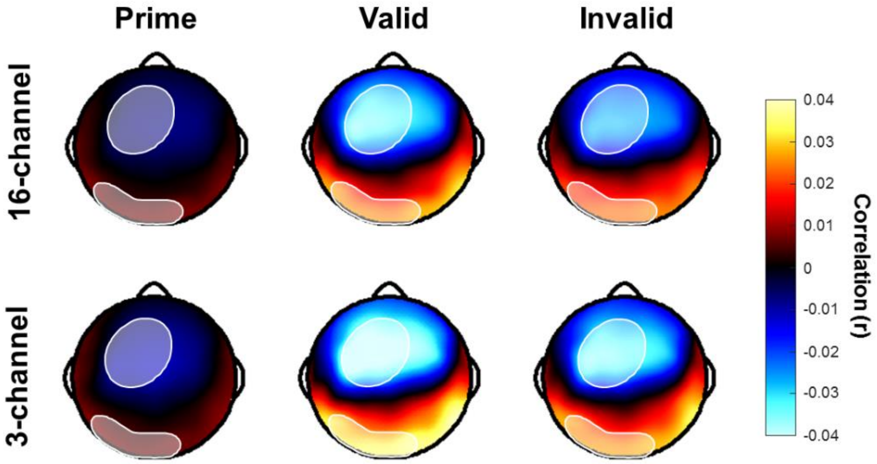


Figure 2.3: Scalp topographies for the latency window of interest (25 - 125 ms). EEG was recorded with a 128-channel electrode cap, and sampled at 1000 Hz with an online average reference. The recordings were then bandpass filtered from 1 to 50 Hz. Thirty channels that showed the largest (regardless of polarity) cross-correlations were selected from the grand average for each of the six conditions, and were combined to form an ROI of 44 channels, indicated by the shaded region.

To estimate the variability in these means, a bootstrap simulation was performed. A control distribution was constructed by replacing the sentences on each trial with random sentences not presented on that trial, and performing the same analysis described above. This control was useful because it maintained all of the average spectral and temporal characteristics of the experimental sentences but was unrelated to the sentences used on a particular trial. Therefore, any nonzero values in the control cross-correlations were due purely to chance. For each subject, this bootstrap was repeated 10 times, and grand-averaged means were computed 10,000 times from these bootstraps to create the control distribution. Mean values were considered significantly non-zero if they fell outside the 99th or 1st percentiles of this distribution.

## RESULTS

### *Behavior*

To ensure that our intelligibility manipulations had their intended effect, we gathered subjective intelligibility ratings prior to each recording block. Participants were asked to indicate on a 1-6 scale how intelligible the target vocoded sentence was. On valid prime trials (when the target and prime matched), participants were asked not to consider the prime when making their intelligibility judgements. Shown in Figure 2.4a, median intelligibility ratings across subjects for the 16-channel blocks were almost exclusively 6, the main exception being the invalid prime condition in the first block. Presumably, this reflects the fact that even highly intelligible vocoded speech can seem foreign when heard for the very first time, and we only gathered a single intelligibility rating for each prime condition (valid/invalid) prior to 16-channel blocks. Median intelligibility ratings for the 3-channel blocks demonstrate a consistent difference between valid and invalid prime conditions across all four blocks. Median intelligibility ratings for the valid prime trials range between 4.5 and 5, while median intelligibility ratings for the invalid prime trials range between 1 and 2.

During experimental trials, participants were asked to indicate whether or not probe clips of vocoded material were drawn from the vocoded target (Figure 2.1). The duration of these probes were adaptively varied according to a 2-down/1-up tracking procedure, and estimated durations required for 71% correct are shown in Figure 2.4b. For ease of discussion, these estimated durations will be referred to as duration thresholds. For the 16-channel vocoded blocks, probe duration thresholds are virtually the same for valid and invalid trials, with a mean duration across blocks of 174 ms for valid trials and 186 ms for invalid trials. This is consistent with intelligibility ratings that are also similar across valid and invalid trials. For the 3-channel vocoded blocks, probe thresholds are substantially higher for invalid trials than for valid trials, with a mean duration across blocks of 1043 ms for invalid trials, and 612 ms for valid trials. This is again consistent with intelligibility ratings that show substantial differences between valid and invalid trials.

As expected, these behavioral results suggest that intelligibility depends on the number of vocoded channels, and indicate no evidence of practice effects on the intelligibility of vocoded speech over the course of the experiment. Specifically, 3-channel vocoded speech with invalid primes remained consistently unintelligible over the course of the experiment.

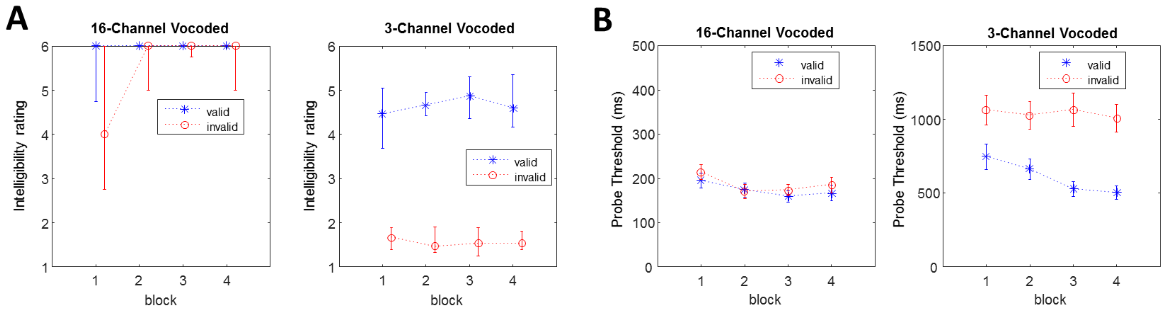


Figure 2.4: Behavioral results. (A) Median subjective intelligibility ratings (25th and 75th quartiles shown) measured before each block. (B) Threshold probe durations for each block (note change in scale).

*Electrophysiology*

A latency range for the N1 peak was defined (25-125 ms) based on grand-averaged data, and a subset of maximally-responding channels were selected to form a region of interest (ROI) within this latency range (Figure 2.2). Cross-correlation means were then selected from the ROI time series, and a bootstrap was performed to estimate a noise floor for cross-correlation means due to chance (Figure 2.5). A 2-factor repeated measures ANOVA on the cross-correlation means failed to reveal a significant interaction between cue validity and number of vocoded channels $F(1, 12) = 0.14$, $p = 0.714$. Consistent with Millman et al. (2015) then, we failed to demonstrate a significant effect of intelligibility on the entrainment response. The ANOVA revealed a significant main effect of cue validity (valid vs. invalid): $F(1, 12) = 13.6$, $p = 0.03$, suggesting that prior knowledge leads to an increased entrainment response. The ANOVA failed to reveal a significant main effect of number of vocoded channels (16 vs. 3): $F(1, 12) = 2.7$, $p = 0.126$, suggesting that sensory detail, at least with regard to 16- and 3-channel vocoded speech, has no effect on the entrainment response. However, as the entrainment response to 3-channel vocoded speech is larger across all conditions than the entrainment response to 16-channel speech, the lack of significant main effect we report may reflect of a lack of statistical power.

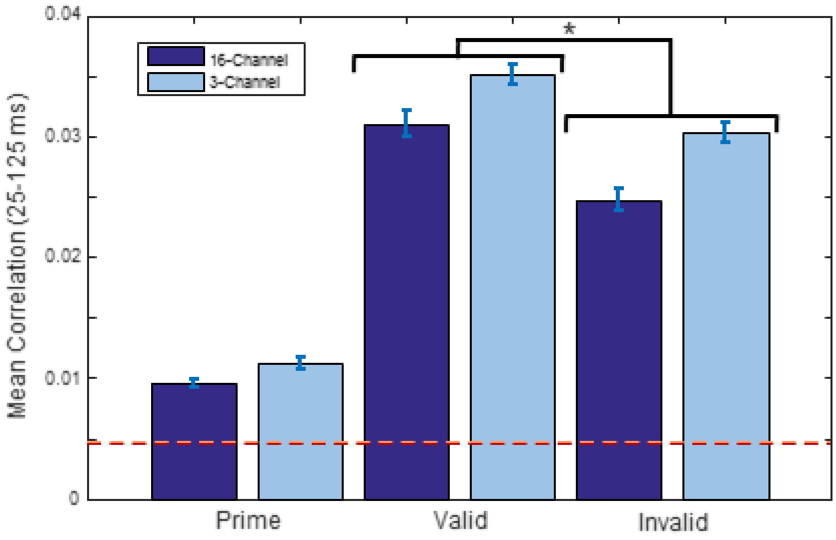


Figure 2.5: Bar plots showing mean correlation values over 25-125 ms latency window and averaged across subjects. The bootstrapped noise floors (99th percentile cutoffs) are displayed here as dotted lines (16-channel = red/dark and 3-channel = pink/light). A 2-factor repeated measures ANOVA revealed a significant effect of Valid vs. Invalid: $F(1, 12) = 13.6$, $p = 0.03$; but no significant effect of 16- vs. 3-channel vocoding: $F(1, 12) = 2.7$, $p = 0.126$; and no significant interaction: $F(1, 12) = 0.14$, $p = 0.714$. Post-hoc t-tests revealed a robust difference between the natural-speech cue and the vocoded targets (all $p < .005$). This suggests that task demand may play a crucial role in determining the strength of the entrainment response.

When we compare the entrainment response to the vocoded targets and the entrainment response to the natural-speech primes, we observe a significant difference. Post-hoc t-tests revealed a significant difference between the natural-speech cue and the vocoded targets (all $p < .005$, with a Bonferroni-corrected $p$-value of .0125) such that the entrainment response to the natural-speech cue is significantly smaller than the entrainment response to vocoded targets. Effect sizes for all comparisons are greater than 1.7, suggesting that this effect is particularly robust. We chose to run a post-hoc analysis on the clean-speech primes rather than include them in the ANOVA for two reasons. First, the clean-speech primes were not directly probed with a behavioral task, and were therefore subjected to qualitatively different task demands than the vocoded speech. Second, our ANOVA was meant to address the effects of priming, sensory detail, and intelligibility, and as such we decided that the analysis of the clean-speech primes should be post-hoc.

## DISCUSSION

The data reported here suggest that (1) the strength of the entrainment response to speech depends on prior knowledge, (2) the strength of the entrainment response to speech seems to critically depend on task demand (prime vs target), and (3) the strength of the entrainment response to speech does not depend on intelligibility, although a lack of significance should not necessarily be overemphasized, especially with a modest sample size (n = 13).

### *Prior acoustic knowledge modulates the entrainment response*

We found that for both the 16-channel and 3-channel vocoded speech, the entrainment response in the valid prime condition is larger than the entrainment response in the invalid prime condition. This suggests that prior knowledge modulates the entrainment response. Furthermore, we did not find a significant difference between the entrainment response to 3-channel and 16-channel vocoded speech, which is to say we did not observe an effect of sensory detail. These results are consistent with the results of Sohoglu et al. (2012), who found an effect of prior knowledge while failing to find an effect of sensory detail on the EEG response. However, while Sohoglu et al. (2012) found the most robust effects of prior knowledge in the 270-700 ms post-onset latency range, we did not observe a reliable entrainment response at these later latencies, and our analysis was restricted to the 25-125 ms latency range. Nonetheless, the effects of prior knowledge and sensory detail on the entrainment response are broadly consistent with previously reported effects of prior knowledge and sensory detail on the auditory evoked response.

A lack of significant interaction between prior knowledge and sensory detail suggest that the effect of prior knowledge is primarily acoustic. Since 16-channel speech is more intelligible than 3-channel speech, we would expect the effect of prior knowledge to be more substantial for 16-channel speech if we suppose that linguistic prior knowledge is modulating the entrainment response. However, if acoustic prior knowledge is modulating the entrainment response, we would expect no difference in the effect of prior knowledge across vocoding conditions, and this is what we observe.

*Task demand modulates the entrainment response*

While we did not set out to explicitly test the effect of task demand on the entrainment response, our results suggest that the strength of the entrainment response depends on the task-relevance of the stimulus. Specifically, we found that the entrainment response to the natural-speech prime sentences were far smaller than the entrainment response to vocoded targets, regardless of condition. Considering that natural speech contains no artificial acoustic distortions, it seems unlikely that stimulus acoustics alone explain this observation. However, Ding et al. (2014) found that in the 1-4 Hz range, the entrainment response to 4-channel vocoded speech was significantly larger than the entrainment response to natural speech, suggesting that even in the absence of unequal task demands, we might expect the entrainment response to natural speech to be smaller than the entrainment response to vocoded speech. Speech information is carried strictly in the envelope for vocoded speech, while for natural speech, there is information in the fine structure that may cause listeners to deploy less resources to the encoding of the envelope, and could therefore help explain why the entrainment response is smaller to the natural-speech cue than to the vocoded target. In Figure 2.5 however, we see that the entrainment response to the natural-speech primes are substantially reduced relative to the vocoded targets, beyond which is expected from the results of Ding et al. (2014). This suggests that the difference in stimuli alone is not sufficient to explain our result. Instead, it suggests that task demand may play a crucial role in determining the strength of the entrainment response, reinforcing the importance of adequately controlling behavioral tasks across intelligibility conditions.

In other words, the fact that listening closely to the natural-speech primes was not necessary to perform the behavioral task may be responsible for the difference in entrainment to the natural-speech primes and the vocoded targets. Participants were asked to compare probe vocoded clips to the target vocoded sentence, which requires acute attention to the vocoded target. On the other hand, while listening to the natural-speech prime affects the intelligibility of the vocoded target on valid trials, acute attention is not required. While it is well-known that attention modulates the entrainment response in multi-talker listening scenarios (e.g. Ding & Simon, 2012; Horton et al., 2013), our results suggest that attention is crucial even in the absence of a competing talker. However, given that we did not explicitly control for task demand across listening to the natural-speech primes and vocoded targets, further studies that explicitly control for and quantify attentional demands are needed.

*Intelligibility does not modulate the entrainment response*

Following Millman et al. (2015), we used priming to study the effect of intelligibility on the entrainment response to vocoded speech, and consistent with Millman et al. (2015), we failed to find an effect of intelligibility. Crucially, we failed to find an effect of intelligibility using novel speech stimuli on each trial and an attentionally demanding behavioral task that was adaptively varied to achieve equal percent correct across conditions. Thus, our result is consistent with previous results suggesting that entrainment to naturalistic speech envelopes is driven by acoustic rather than linguistic neural processes (Howard & Poeppel, 2010; Millman et al., 2015; Zoefel & VanRullen, 2016).

A potential confound in our study however, is the fact that our behavioral task explicitly directed listeners to the acoustics of the speech stimulus, and it is possible that if we used a task that more explicitly focused on the semantic and/or syntactic aspects of the stimulus, we may have observed an effect of intelligibility. However, probe duration thresholds are shortest for 16-channel vocoded speech, are longer for validly-primed 3-channel vocoded speech, and are still longer for invalidly-primed 3-channel speech, thus neatly scaling with subjective intelligibility reports (Figure 2.4). This suggests that listeners are making use of linguistic rather than strictly acoustic information when comparing the probe to the target, at least to the extent that "linguistic" refers to those over-learned features of the clean-speech stimulus that allow for the restoration of intelligibility to degraded speech.

Another potential confound is the fact that the effect of prior knowledge might not be equal across vocoding conditions. Since 16-channel vocoded speech more closely resembles natural speech than 3-channel vocoded speech, we might expect the effectiveness of the prime to be greater for 16-channel speech. If the prime was more effective at modulating the entrainment response in the 16-channel vocoded condition, it could mask a subtle effect of intelligibility in the 3-channel condition. If validly-primed 3-

channel vocoded speech is modulated by both prior knowledge and intelligibility, and validly-primed 16-channel vocoded speech is modulated only by prior knowledge, we would expect to see a significant interaction between prior knowledge and sensory detail. However, if prior knowledge is disproportionately effective at modulating the entrainment response to 16-channel vocoded speech, we may fail to observe this interaction. Nonetheless, our data are consistent with literature suggesting that the entrainment response is primarily driven by acoustic properties of the speech signal.

One hypothesis, put forward by Doelling et al. (2014), is that entrainment in the theta band is necessary but not sufficient for speech perception. This hypothesis is based on the finding that the magnitude of abrupt changes at the start of syllables are predictive of syllable length (Greenberg et al., 2003). It is possible then, that neural entrainment is driven by this initial syllabic chunking mechanism, which is acoustic in origin but may help support subsequent linguistic processes.

### *Interactions among multiple sources*

It is almost certainly the case that the scalp-recorded entrainment response we report here is the resulting sum of multiple underlying sources that are driven by different aspects of the speech stimulus. However, since we could not resolve distinct sources that were consistent across subjects in this dataset, we report only the sum of these sources. It is therefore possible that sources driven by acoustic properties of the speech are masking a source that is modulated by intelligibility.

### *Conclusions*

While the strength of the entrainment response to speech depends on prior knowledge, we failed to find evidence that the entrainment response depends on intelligibility. However, a number of potential confounds make it difficult to rule out that our design masked a subtle effect of intelligibility. Finally, the strength of the entrainment response to speech seems to critically depend on task demand, underscoring the importance of controlling task demand in entrainment studies.

# CHAPTER 3

## INTRODUCTION

The ability to parse acoustic stimuli at multiple timescales is an important feature of the human auditory system that allows us to understand hierarchically organized sound structures, such as those central to speech and music. For speech, it has been suggested that coupled cortical oscillations entrain to fluctuations in speech energy/information to support the parsing of the signal into discrete linguistic units (e.g. Ghitza, 2011; Giraud & Poeppel, 2012; Canolty et al. 2006; Ding et al., 2016; Wilsch et al., 2018). For music, it has been suggested that coupled oscillations support hierarchical beat perception and sensorimotor synchronization (e.g. Large & Palmer, 2002; Repp, 2005; Fujioka et al., 2012; Nozaradan et al., 2012). While speech and music share certain temporal properties (Ding et al., 2017), the extent to which the functional roles of these oscillations are shared across speech and music though, is not known, nor is the extent to which these oscillations play a general role in the perception of temporal regularities.

Ding et al. (2016) identified a cortical response that entrains to periodic fluctuations in high-level linguistic units (phrases and sentences) in the absence of acoustic cues, providing evidence not only that cortical responses can entrain to linguistic-informational rather than acoustic periodicities, but that speech is processed simultaneously at multiple time-scales. Nozaradan et al. (2011) found that when subjects were asked to subdivide a sequence of pure tones into discrete units of two or three tones, a cortical response could be seen at the frequency of the subdivision, thereby identifying a cortical response that entrains to the subject’s perception of meter. Nozaradan et al. (2012) found a similar result in the absence of explicit metric grouping instructions, suggesting that the cortical response also entrains to the spontaneous perception of meter.

The primary goal of the present study was to investigate whether or not hierarchically organized musical structures are parsed simultaneously at multiple time-scales, as indicated by cortical entrainment responses to hierarchically organized musical sequences. A secondary goal was to extend the findings of Ding et al. (2016), who reported an entrainment response to phrasal and sentential sequences, to word-level and semantic-level sequences.

## METHODS

### *Participants*

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Twenty English-speaking listeners (11 female; age range: 18-61; mean age: 26) with normal hearing participated in the study. Twelve of these listeners were musically trained, and the other eight had no musical training. Of these twelve listeners with musical training, eight had formal theory training. No listeners self-reported having absolute pitch.

### *Stimuli*

Listeners were presented with three different stimulus types: sentences, words, and triads. Sentences consisted of a three-word noun phrase followed by a three-word verb phrase to form a six-word sentence (e.g. The long fight caused much harm), and were largely based on sentence materials used by Ding et al. (2016). Words consisted of three distinct syllables that combine to form a single word (e.g. com - pa - ny). Triads consisted of three nonsense syllables (/do/, /beɪ/, or /lɑ/) with pitches corresponding to the three notes of a major triad, in ascending order (e.g. 220 Hz - 277.2 Hz - 329.6 Hz, or A3 - C#4 - E4) All stimuli were generated using a Klatt text-to-speech synthesizer in Pratt (Version 6.0.36). All stimuli were composed of individual sound units that were presented at a rate of 3 Hz. For sentence stimuli, these sound units were single words, for word stimuli, these sound units were single syllables, and for triad stimuli, these sound units were nonsense CV syllables with specific pitches.

Figure 3.1 shows the hierarchical structure of the three stimulus types. For sentence stimuli, single words are presented at 3 Hz, individual linguistic phrases are presented at 1 Hz, and sentences are presented at 0.5 Hz (Figure 3.1a). For word stimuli, individual syllables are presented at 3 Hz, and whole (3-syllable) words are presented at 1 Hz (Figure 3.1b). For triad stimuli, specific-pitch nonsense syllables were presented at 3 Hz, individual triads were presented at 1 Hz, and harmonically-related pairs of triads were presented at 0.5 Hz (Figure 3.1c). Specifically, the second chord is a dominant (or V) chord relative to the first chord, which is always the tonic (or I) chord.

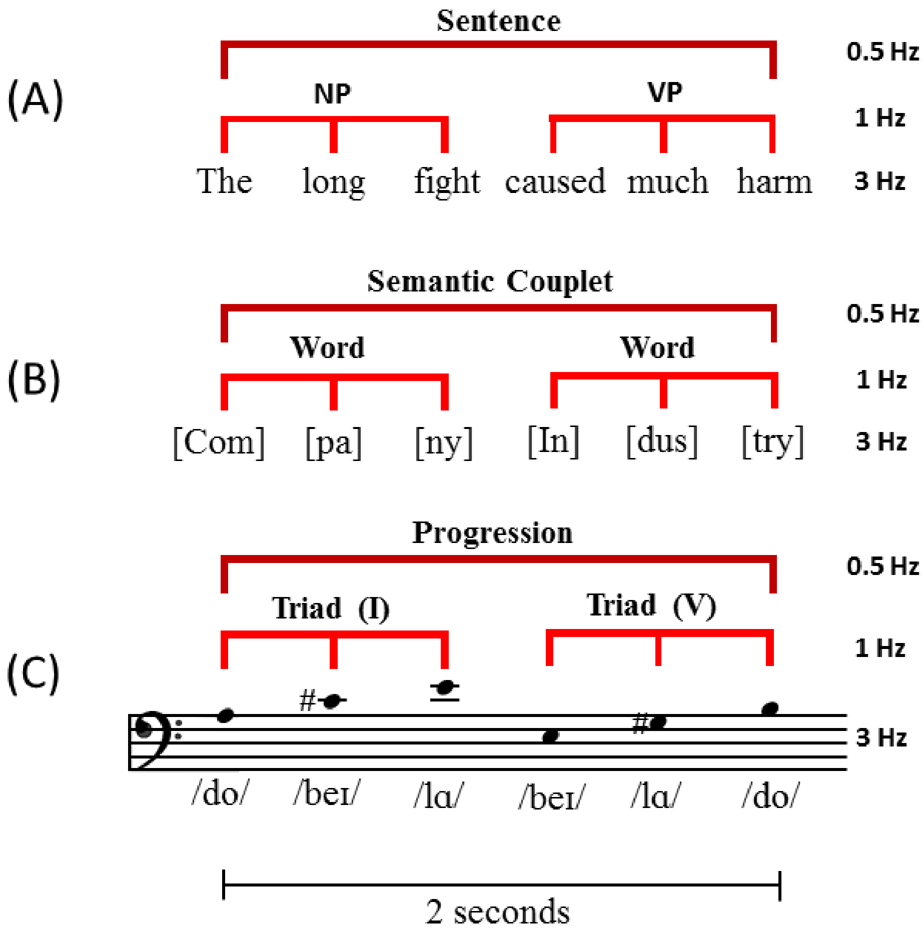


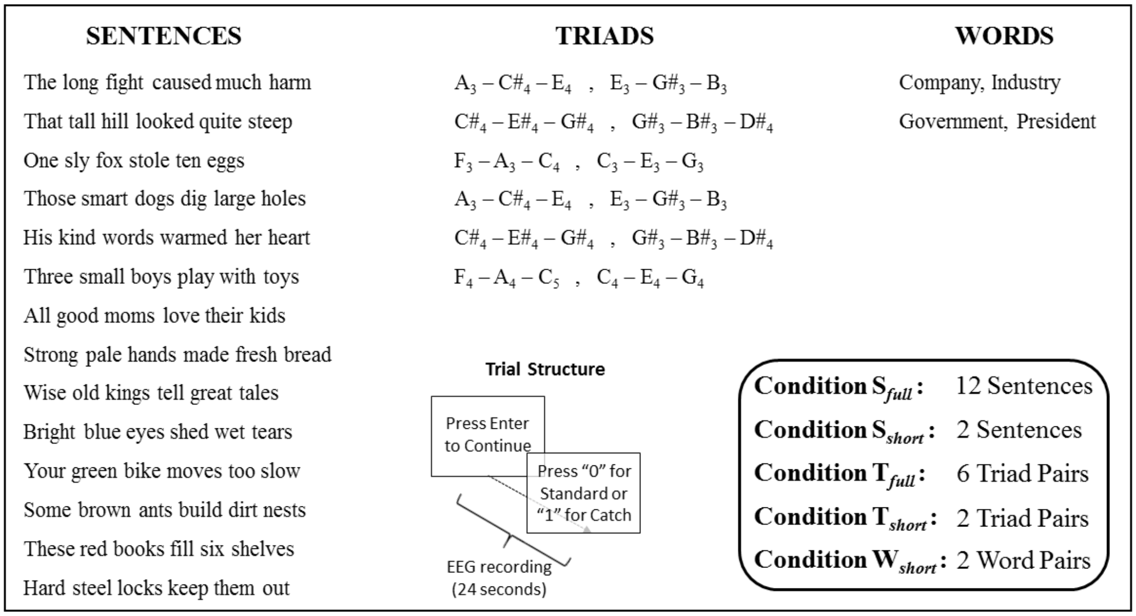
Figure 3.1: Schematic of different stimulus types. (**A**) Individual words are presented at 3 Hz, and form a distinct noun phrase (NP) and verb phrase (VP) at 1 Hz. These 1-Hz phrases combine to form a complete sentence at 0.5 Hz. (**B**) Individual syllables are presented at 3 Hz, and form distinct words at 1 Hz. These two words are semantically related, and combine to form a semantic couplet at 0.5 Hz. (**C**) Individual sung syllables are presented at 3 Hz, and form distinct ascending triads at 1 Hz. These two triads combine to form a I-V chord progression at 0.5 Hz.

Although there were only three stimulus types, there were a total of five experimental conditions (Table 3.1). In condition $S_{full}$, 12 unique sentences were selected, and on each trial, were presented in a pseudo-random (circularly shifted) order. Specifically, the initial order of the 12 sentences was random, and for each trial, the starting sentence was circularly shifted by a random number of sentences. This pseudorandomization was done in order to match the triad condition (condition 3), described in the subsequent paragraph. In condition $S_{short}$, 2 unique sentences were selected, and on each trial, were presented in a random alternating order (either ABAB... or BABA…). A total of 14 sentences were generated for this experiment, and are listed in Table 3.1. For each listener, 12 of these sentences were randomly assigned to condition 1, and the remaining two sentences were assigned to condition 2. Following Ding et al. (2016), condition $S_{full}$ utilizes a large stimulus set such that we do not expect effects of stimulus overexposure. Condition $S_{short}$ however, utilizes a limited stimulus set, allowing us to determine whether stimulus novelty has an effect on the cortical tracking response.

In condition $T_{full}$, a series of six tonic-dominant triad pairs were constructed (Table 3.1). The pitch class of each new pair was shifted (a shift in pitch class is equivalent here to a change in key signature) up by a major third relative to the previous pair, which means that the pitch classes of the triad repeat every third pair. However, we shifted the octave of the sixth pair (relative to the third pair) so that the repeating sequence was independent of pitch height. Finally, two copies of these 6-pair sequences were concatenated to yield a 12-pair sequence. On each trial, the starting pair was circularly shifted by a random number of pairs so that the sequence was novel but the structure of the pitch class shift was preserved. Since absolute pitch is extremely rare, even among highly-trained musicians (Deutsch et al., 2006), we did not control for repeating pairs over the course of a trial. Furthermore, we chose a fixed-interval pitch class shift instead of a random-interval pitch class shift in order to avoid jarring or surprising transitions between pairs. In condition $T_{short}$, two adjacent triad pairs were selected from the six available triad pairs and presented in random alternating order (either ABAB... or BABA…). For both conditions 3 and 4, syllables were pseudo-randomized to maximize the distance between repetitions of a single syllable.

In condition W*short*, two unique word pairs were generated (Table 3.1). These four words were selected based on their high frequency of use and on the fact that they are content rather than function words. Furthermore, they formed semantically similar pairs. On each trial, these word pairs were presented in a random alternating order.

Table 3.1: List of stimuli used. A total of fourteen sentences were generated, with no words repeating. A sequence of six chord progressions was generated (subscript numbers refer to octave designation). Two word pairs were generated.



| SENTENCES | TRIADS | WORDS |
|---|---|---|
| The long fight caused much harm | $A_3 – C\#_4 – E_4$ , $E_3 – G\#_3 – B_3$ | Company, Industry |
| That tall hill looked quite steep | $C\#_4 – E\#_4 – G\#_4$ , $G\#_3 – B\#_3 – D\#_4$ | Government, President |
| One sly fox stole ten eggs | $F_3 – A_3 – C_4$ , $C_3 – E_3 – G_3$ | |
| Those smart dogs dig large holes | $A_3 – C\#_4 – E_4$ , $E_3 – G\#_3 – B_3$ | |
| His kind words warmed her heart | $C\#_4 – E\#_4 – G\#_4$ , $G\#_3 – B\#_3 – D\#_4$ | |
| Three small boys play with toys | $F_4 – A_4 – C_5$ , $C_4 – E_4 – G_4$ | |
| All good moms love their kids | | |
| Strong pale hands made fresh bread | | |
| Wise old kings tell great tales | | |
| Bright blue eyes shed wet tears | | |
| Your green bike moves too slow | | |
| Some brown ants build dirt nests | | |
| These red books fill six shelves | | |
| Hard steel locks keep them out | | |

Trial Structure
Press Enter to Continue
Press “0” for Standard or “1” for Catch
EEG recording (24 seconds)

| | |
|---|---|
| **Condition $S_{full}$:** | 12 Sentences |
| **Condition $S_{short}$:** | 2 Sentences |
| **Condition $T_{full}$:** | 6 Triad Pairs |
| **Condition $T_{short}$:** | 2 Triad Pairs |
| **Condition $W_{short}$:** | 2 Word Pairs |

*Task*

Each participant sat facing a computer monitor in a single-walled sound-attenuated booth with sound-treated interior walls. Stimuli were presented diotically at 70 dB SPL over Final 500 electrostatic loudspeakers positioned at 45˚ and -45˚ degrees relative to the listener, at a distance of five feet from the listener’s head. There were five experimental blocks, corresponding to five experimental conditions, and the order of these blocks were randomized for each listener. Each block contained 25 trials. Each trial had a duration of 24 seconds, and contained 12 individual stimuli (sentences, semantic couplets, or progressions, see Figure 3.1). All five blocks were tested in a single session. Following Ding et al. (2016), listeners were instructed to detect catch trials. On a catch trial, the individual sound elements of a sentence/word/triad were reversed in order (e.g. that tall hill looked quite steep → steep quite looked hill tall that, [in]-[dus]-[try] → [try]-[dus]-[in], $E_3$-$G\#_3$-$B_3$ → $B_3$-$G\#_3$-$E_3$, etc.). For sentence stimuli, a catch trial contained a single reversed sentence, while for word and triad stimuli, a catch trial contained two reversed words and two reversed triads, respectively. Each block contained five catch trials, which were excluded from the EEG analysis.

Prior to testing, listeners were asked to familiarize themselves with the stimuli on a self-paced GUI. As many of our listeners were unfamiliar with synthesized speech, we asked them to listen to the different sentences and words (which were accompanied by text on the screen) until they were convinced they could understand the speech material without the aid of text.

*EEG Recording and Pre-Processing*

High-density EEG (128 channels) was recorded with equipment from Neuroscan. Electrodes were placed following the international 10/5 system (Oostenveld & Praamstra, 2001), and all channel impedances were kept below 10 kΩ. The EEG data was sampled at 1000 Hz, and filtered offline with a passband of .15 to 50 Hz. The filtered data were then segmented into individual trials which were 22 seconds long, beginning 2 seconds after the start of the sentence (yielding a frequency bin size of 1/22 Hz). This delay was incorporated to remove the onset response to the start of the stimulus. Artifacts were removed from the segmented EEG data using the Fast ICA algorithm (Hyvarinen & Oja, 1997).

Following Ding et al. (2016), the EEG responses were then denoised using the Denoising Source Separation (DSS) algorithm, which is a blind source separation technique that extracts neural response components that are consistent across trials (de Cheveigne & Simon, 2008). DSS computes a bias function based on the averaged neural data, and applies a transformation to the unbiased (raw) neural data that maximizes this bias function. This bias function was computed across (rather than within) experimental conditions to avoid the artificial introduction of differences across these conditions.

### *Analysis*

The denoised EEG data were analyzed in the frequency domain. For each subject, an average was taken over trials, and a Discrete Fourier Transform (DFT) was applied to the averaged data. In order to remove the 1/f trend in the denoised data, the magnitude of each Fourier coefficient was normalized by the median magnitude of neighboring coefficients while the phase spectrum was left intact. Coefficient magnitudes at 0.5 Hz and above were normalized by the median magnitude of the coefficients at ± 11 bins (1/2 Hz). In order to maintain symmetry about the coefficient being normalized, coefficient magnitudes between 0.25 and 0.5 Hz were normalized by the median magnitude of the coefficients at ± 6 bins (1/4 Hz). Coefficients below 0.25 Hz were discarded. The median was used so that sharp peaks were not artificially boosted by the normalization procedure. Since the median ignores outlier values, the normalized magnitude of coefficients neighboring sharp peaks were not affected by the peaks themselves.

In order to find optimal channel weights for each normalized frequency bin, a Singular Value Decomposition (SVD) was applied to local portions of the normalized spectrum. A local portion was defined as 9 bins centered on the frequency bin of interest (± 1/6 Hz). The local portion around each frequency bin was submitted to an SVD, and the first component was selected. The first component of the SVD reveals the spectrum that captures the most variance, along with the associated channel weights. These channel weights refer to the relative distribution of activity across channels for a local frequency region. The magnitude of the Fourier coefficient at the bin of interest was divided by the median of the magnitudes of the Fourier coefficients on either side (± 1/6 Hz). By converting the absolute magnitude of the bin of interest to an SNR measurement, we can correct for unequal variance across SVD components for different portions of the spectrum.

### *Statistical Analysis*

This SVD procedure yielded a spectrum of optimally-weighted magnitudes, and statistical analysis was performed on this optimized spectrum. One-tailed, two-sample t-tests were used to test whether the cortical response in a frequency bin was significantly stronger than the average of the neighboring four frequency bins (two bins on either side). T-tests were performed for each frequency bin between 0.25 and 3.5 Hz, and an FDR correction for multiple comparisons (Benjamini & Hochberg, 1995) was applied ($\alpha = 0.01$).

## RESULTS

### *Behavior*

To ensure that listeners were paying attention to the stimuli, 1/5 of the trials in each block were catch trials, and after each trial, listeners indicated whether they thought the trial was a standard or catch trial (see *Task* in METHODS for details). Percent correct for standard and catch trials are shown for each condition in Figure 3.2. In general, listeners identified standard and catch trials above chance and with a high degree of accuracy. Chance was computed based on a probability matching observer, who randomly indicates standard on 20/25 trials and catch on 5/25 trials. Such an observer would correctly identify 80% of standard trials, and would correctly identify 20% of catch trials. It was somewhat surprising that performance for condition $S_{short}$ was worse than condition $S_{full}$, since condition $S_{short}$ offered a more restrictive stimulus set, and suggests that listeners may not have been attending as closely to the stimuli in $S_{short}$ relative to $S_{full}$. Due to the differences in stimuli across conditions, and the fact that an adaptive procedure was not used to equate performance across conditions, a statistical comparison of the behavioral data across conditions was not pursued.

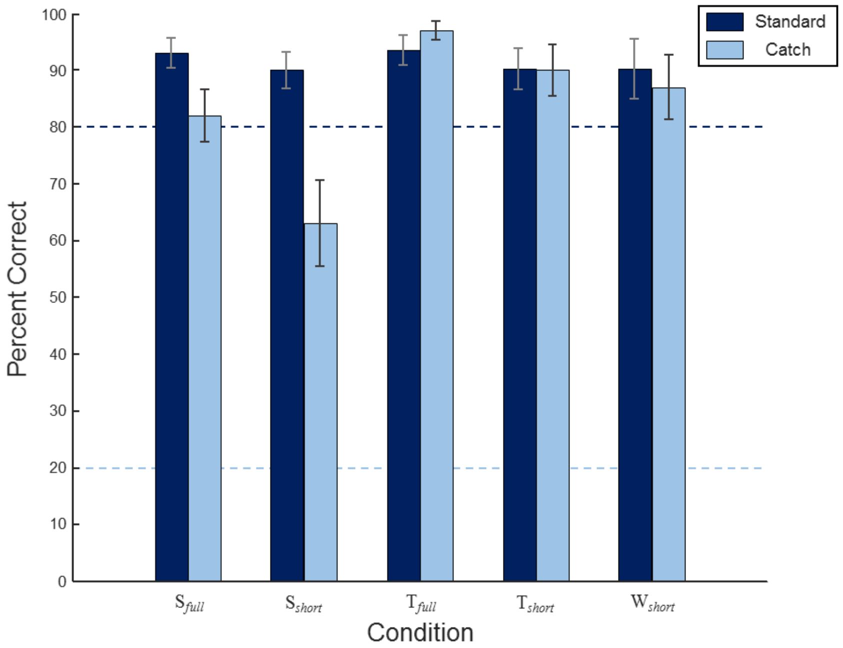


Figure 3.2: Percent correct for each experimental condition, shown separately for standard and catch trials. Chance performance based on a probability-matching observer for standard (dark blue) and catch (light blue) trials are indicated with dotted lines.

*Stimulus Envelopes*

Across conditions, 3-Hz sound sequences were organized in a hierarchical fashion, forming 1-Hz and 0.5-Hz sequences of linguistic/musical units. Since cortical envelope tracking in the delta/theta range is known to be robust (Ding & Simon, 2012; Horton et al., 2013; Doelling et al. 2014), a potential concern is that 1-Hz and 0.5-Hz energy in the stimulus envelope could generate a cortical response at these frequencies. For each condition, 500 example trial stimuli (24-second duration) were selected, a Hilbert transform was applied, and the absolute value was taken to extract the envelope from each example stimulus. The DFT of the stimulus envelope of each stimulus was taken, and magnitudes were averaged over examples, yielding average stimulus modulation spectra for each condition, shown in Figure 3.3.

The modulation spectra provide predictions for the auditory envelope-tracking response. Across conditions there is a large peak at 3 Hz, which corresponds to the rate at which sounds were presented (Figure 3.1). Also, the modulation spectra of $S_{full}$ and $T_{full}$ are flatter, respectively, than $S_{short}$ and $T_{short}$. This is because the longer the overall period, the smaller the fundamental frequency, and energy at the sidebands of the 3-Hz peak will be distributed to a greater number of bins. For the *short* conditions, there is noticeable energy at 1 Hz and 0.5 Hz, but peaks at these frequencies are not obviously larger than peaks at other frequencies.

Overall, the energy in the envelope of the stimulus at 1 Hz and 0.5 Hz is not greater than the energy at other peaks. This suggests that if significant energy in the cortical response is observed at these two frequencies, but not at other frequencies with greater or similar energy in the stimulus envelope, these responses cannot be interpreted as reflecting the envelope-tracking response.

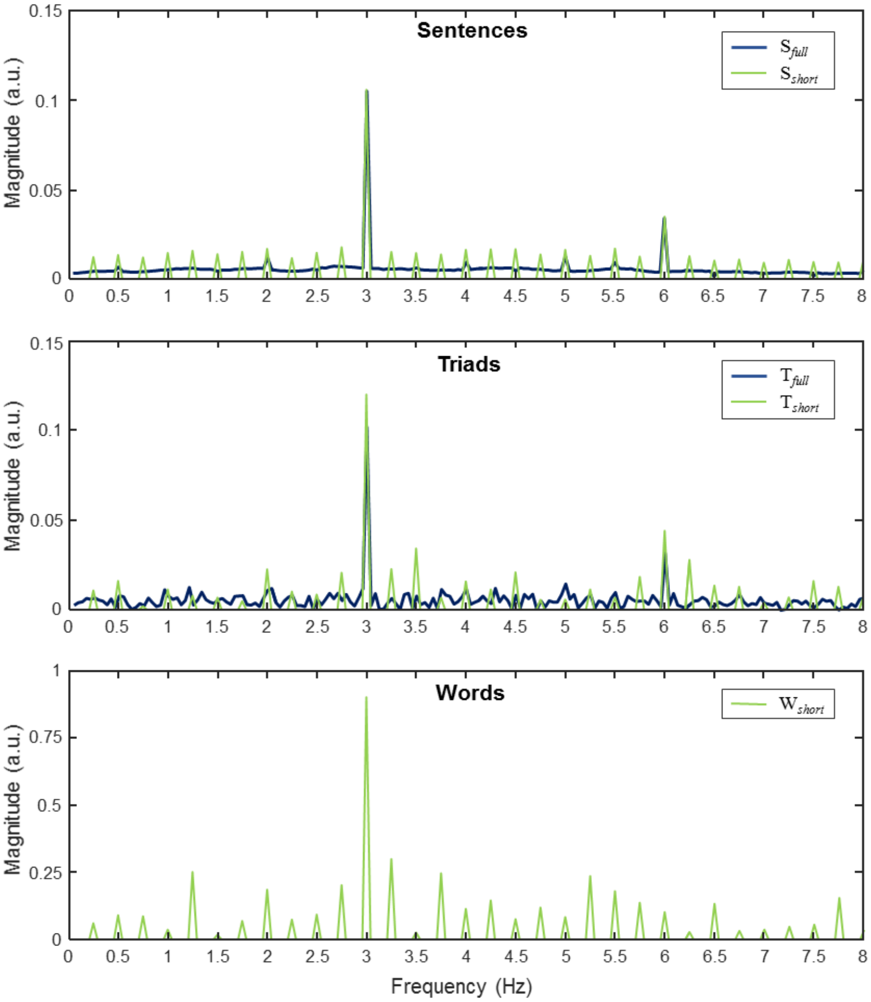


Figure 3. 3: Modulation spectra of the stimuli in different conditions. Modulation spectra were computed by taking the DFT of the absolute value of the Hilbert transform of example trial stimuli.

### *Electrophysiology*

We hypothesized that we would observe cortical tracking to the stimulus at three distinct frequencies, corresponding to the hierarchical structure outlined in Figure 3.1. The cortical response spectrum was analyzed from 0.25 to 3.5 Hz, and an FDR correction for multiple comparisons was applied with a cutoff α = 0.01 (yielding a *p*-value criterion of

0.0004). Across conditions, we expect a cortical response at 3 Hz, corresponding to the rate at which sounds were presented. Shown in Figure 3.4a, a significant response at 3 Hz was observed for all conditions. For conditions $S_{full}$ and $S_{short}$, we expect a cortical response at 1 Hz, corresponding to the phrase presentation rate, and at 0.5 Hz, corresponding to the sentence presentation rate. We observed a significant response at both 1 Hz and 0.5 Hz, for both $S_{full}$ and $S_{short}$. For conditions $T_{full}$ and $T_{short}$, we expect a cortical response at 1 Hz, corresponding to the triad presentation rate, and at 0.5 Hz, corresponding to the I-V progression presentation rate. We observed a significant response at both 1 Hz and 0.5 Hz for $T_{full}$, a significant response at 1 Hz for $T_{short}$, and a marginally significant response at 0.5 Hz for $T_{short}$ ($p$ = 0.0019). For condition $W_{short}$, we expect a cortical response at 1 Hz, corresponding to the word presentation rate, and at 0.5 Hz, corresponding to the semantic couplet presentation rate. We observed a significant response at 1 Hz, and a marginally significant response at 0.5 Hz ($p$ = 0.0021).

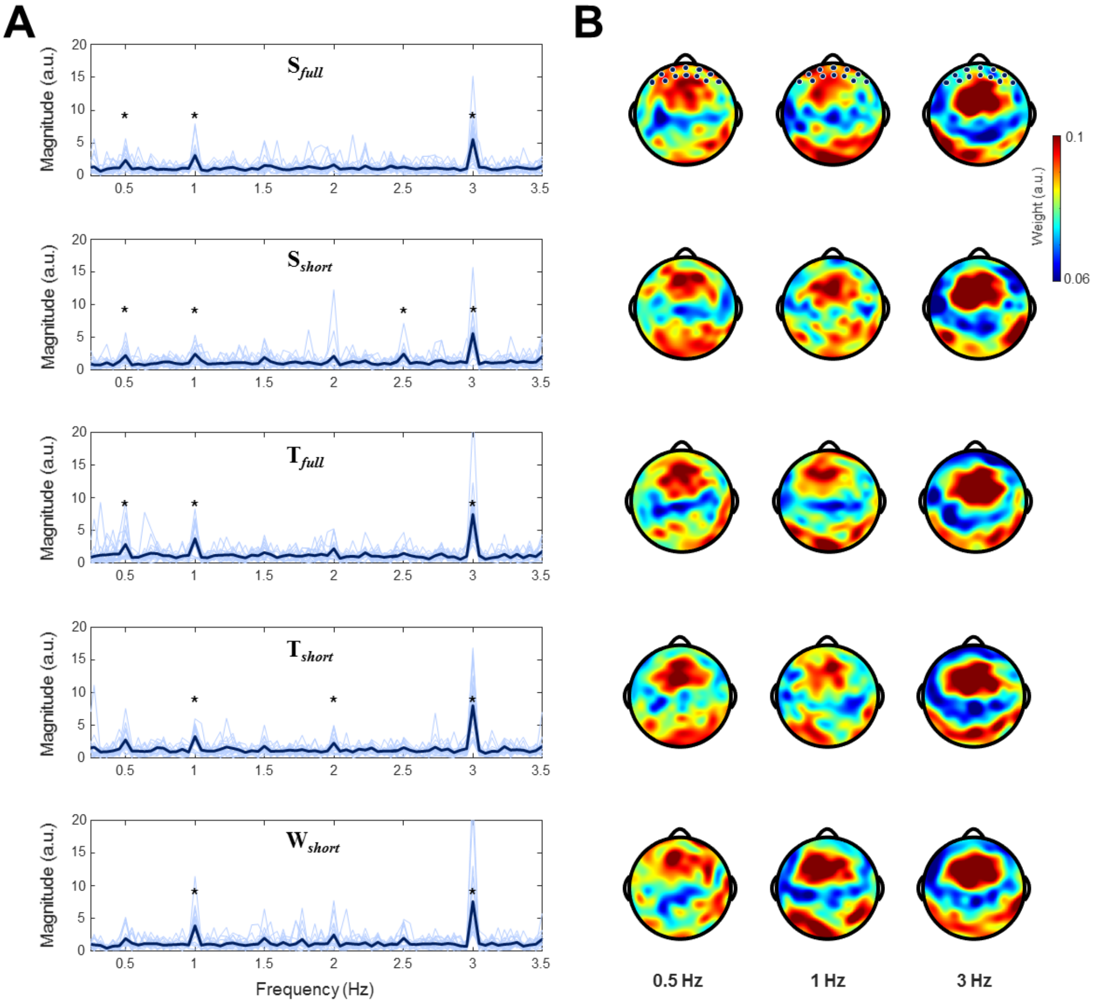


Figure 3.4: (**A**) Cortical response spectra for the different conditions. Individual subjects are shown in light blue traces, and the grand average is shown in dark blue. Cortical tracking of hierarchical sequences in the stimulus are reflected in spectral peaks. Frequency bins with significantly (FDR corrected) more energy than neighboring bins (±2 bins) are marked with an asterisk. (**B**) Scalp-topographic maps of channel weights derived from the SVD procedure. Channel weights can be interpreted as the relative cortical activity across channels associated with each local frequency region centered at 0.5 Hz, 1 Hz, and 3 Hz. Channels included in the frontal ROI are indicated with blue dots corresponding to electrode locations (top row).

We also observe two significant peaks at frequencies that do not correspond to sequences in the stimulus; at 2 Hz in the T*short* condition, and at 2.5 Hz in the S*short* condition. Considering that there isn't substantial energy in the acoustic envelope at these frequencies, it is unlikely that they reflect an envelope-tracking response. Instead, these responses likely reflect distortion products generated by cortical dynamics in response the hierarchical stimulus.

Scalp topographies, shown in Figure 3.4b, are broadly consistent with an auditory envelope-tracking EEG response (Nozaradan et al., 2012; Baltzell et al., 2017). This pattern (fronto-central to posterior dipole) is clearest at 3 Hz, suggesting that an envelope-tracking response may be contributing more strongly to the 3-Hz peak than to the peaks at 0.5 Hz and 1 Hz.

Since frontal sources are more active during the processing of linguistic information as opposed to strictly acoustic information (Demonet et al. 1992; for review, see Friederici, 2002), and since frontal sources are more active during cognitively-controlled as opposed to automatic interval-timing tasks (for review, see Lewis & Miall, 2003), we anticipated differences in relative frontal activity across frequencies. We defined a frontal ROI, (channels included in this ROI are shown in the top row of topographic maps in Figure 3.4b), which was chosen not to overlap with the main dipole to avoid overlap with the envelope-tracking response. We ran a 2-factor (Frequency by Condition) repeated-measures ANOVA to examine the effects of frequency on frontal channel weights. We found a significant main effect of Frequency ($F(2,38) = 7.27$, $p = 0.002$), a significant main effect of Condition ($F(4,76) = 3.67$, $p = 0.009$), and a marginally significant interaction ($F(8,152) = 1.83$, $p = 0.075$). Since the interaction was marginally significant, simple main effects were calculated, yielding a significant effect of Frequency ($F(2,36) = 6.05$, $p = 0.005$) and Condition ($F(4,72) = 3$, $p = 0.024$).

Post-hoc planned comparisons of Frequency revealed a significant (Bonferroni corrected) difference between 0.5 Hz and 3 Hz ($p$ = 0.01) and between 1 Hz and 3 Hz ($p$ = 0.013), but not between 0.5 Hz and 1 Hz ($p$ = 1). Shown in figure 3.5a, frontal activity is significantly higher at 0.5 Hz and 1 Hz than at 3 Hz. Post-hoc comparisons of Condition only revealed a significant difference between $S_{full}$ and $S_{short}$, suggesting that degree of frontal activity in response to speech was not different than in response to music. Since stimulus overexposure may have created unequal attentional demands, we expected a difference between *full* and *short* conditions, and compared frontal activity across *full* ($S_{full}$, $T_{full}$) and *short* ($S_{short}$, $T_{short}$, $W_{short}$) conditions. Shown in Figure 3.5b, this comparison was not significant.

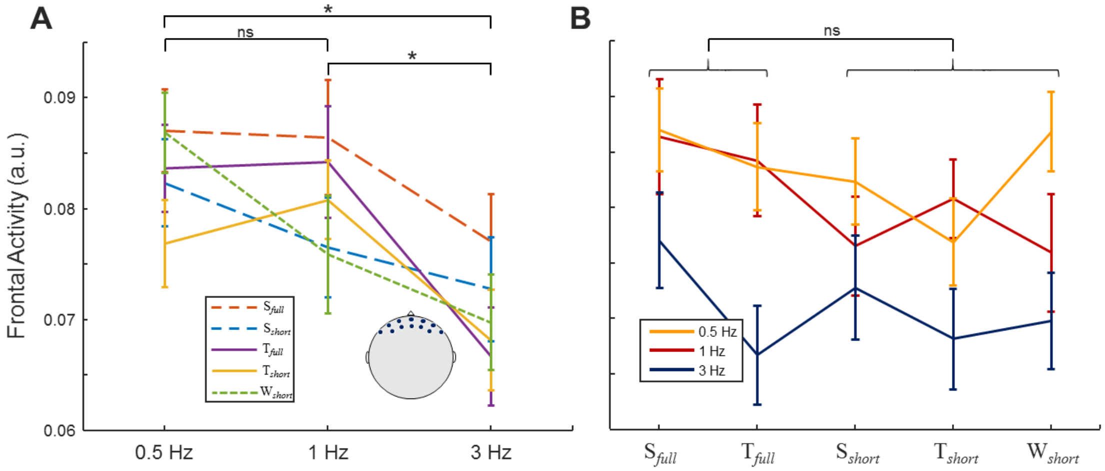


Figure 3.5: (**A**) Relative activity in the frontal ROI (blue dots on gray scalp diagram) for each conditions as a function of frequency. Asterisks indicate significant post-hoc paired comparisons. (**B**) The same data from (A) but shown as a function of condition. The comparison between *full* and *short* conditions was not significant.

*Hemispheric Lateralization*

Despite a largely overlapping architecture for the perception and organization of musical and linguistic sounds (Koelsch et al., 2002; Koelsch et al., 2004), imaging studies have identified hemispheric differences in prefrontal regions recruited for the processing of linguistic and musical stimuli (Zatorre et al., 1992; Zatorre et al, 1994; Zatorre et al., 2002). Specifically, the left prefrontal cortex is more strongly activated by linguistic stimuli, and the right prefrontal cortex is more strongly activated by musical stimuli. To investigate hemispheric lateralization, we split our frontal ROI into left and right subregions (ignoring the midline electrodes), and ran paired t-tests comparing left vs. right in each condition, having collapsed across 0.5 Hz and 1 Hz frequencies. No comparisons were significant (all $p > 0.5$).

*Musicians vs Nonmusicians*

Only 12 out of 20 subjects had musical training, and while we observed a significant response at 0.5 Hz (for $T_{full}$) and at 1 Hz (for $T_{full}$ and $T_{short}$) across all subjects, it is possible that these effects were mainly driven by subjects with musical training. To investigate this, we ran 2-tailed two-sample t-tests between musicians and nonmusicians separately for responses at 0.5 Hz and 1 Hz. These comparisons did not reveal a significant effect of musicianship for responses at 0.5 Hz ($T_{full}$: $p = 0.2$, $T_{short}$: $p = 0.96$) and at 1 Hz ($T_{full}$: $p = 0.81$, $T_{short}$: $p = 0.091$).

## DISCUSSION

Replicating the main result of Ding et al. (2016), we found cortical responses following hierarchically-organized repetitions of linguistic phrases and sentences, in addition to a cortical response following changes in acoustic energy, reiterating that linguistic structures are parsed at multiple timescales. We also identified cortical responses to hierarchically organized musical phrases, suggesting that musical structures, in addition to linguistic structures, are parsed at multiple timescales.

### *Cortical Responses to Speech*

Cortical responses to speech were measured in three conditions: $S_{full}$, $S_{short}$, and $W_{short}$ (Figure 3.1 and Table 3.1). In condition $S_{full}$, using a relatively large set of sentences (n = 12), we found that cortical responses track hierarchically-organized changes in linguistic structure (phrases and sentences). In condition $S_{short}$, using only a pair of sentences, we found a similar result, suggesting that overexposure to the stimulus does not eliminate the cortical response observed in $S_{full}$. In condition $W_{short}$, using a pair of semantically coupled multi-syllabic words, we observed a significant response to the word rate and a marginally significant response to the semantic couplet rate. These results suggest that cortical responses track linguistic-informational boundaries that include words, phrases, and sentences. Importantly, cortical tracking of these boundaries was not reducible to the tracking of acoustic energy. Finally, the marginal significance of the semantic couplet rate suggests that hierarchical organizations of linguistic information might emerge beyond those explicitly specified by the syntactic-level and word-level organizations.

### *Cortical Responses to Music*

Cortical responses to music were measured in two conditions: $T_{full}$, and $T_{short}$ (Figure 3.1 and Table 3.1). In condition $T_{full}$, using a set of triad (chord) progressions that shifted pitch class continuously, we found that cortical responses track hierarchical changes in musical structure. Specifically, we found a significant response to individual triads, and a significant response to pairs of triads with fixed (I-V) harmonic relationships. This suggests that listeners group triads into discrete musical phrases, and group pairs of triads into discrete musical progressions. However, in condition $T_{short}$, we observed only a marginally significant response at 0.5 Hz, suggesting that the cues for grouping progressions are not as robust with a limited set of repeated progressions.

*Frontal activity*

Across conditions and across frequencies, the main dipole resembles a typical auditory envelope-tracking response (Figure 3.4b). However, relative frontal activity is significantly higher at frequencies related to linguistic and musical sequences (0.5 Hz and 1 Hz) than at the 3-Hz stimulus envelope rate (Figure 3.5a). This is consistent with the fact that frontal sources are recruited for the processing of linguistic rather than acoustic information (Freiderici, 2002), and suggests that these sources may be relatively more active at frequencies related to linguistic/musical sequences. However, these frontal regions, in particular the left inferior prefrontal cortex, are involved in both phonological and lexico-semantic processing (e.g. Poldrack et al. 1999), suggesting that frontal activity should be present even at the 3-Hz envelope rate, since individual sounds contained, at the very least, phonological information across conditions. Since we are analyzing the relative rather than the absolute distribution of activity across channels, it is likely that the envelope-tracking response is dominating activity at 3 Hz, potentially to the exclusion of a more linguistic response. Indeed, Ding et al. (2016) report intercranial (EcoG) electrodes in the left frontal cortex that selectively respond to linguistic information at the envelope rate. However, the fact that the scalp topography at 1 Hz in the $W_{short}$ condition strongly resembles the topography at 3 Hz in the $S_{full}$ and $S_{short}$ conditions (Figure 3.4) suggests that the higher-order linguistic computations of phrases and sentences may recruit frontal sources, including working memory circuits, more heavily than lower-order linguistic computation of words (Freiderici et al., 2003; Just et al., 1996).

It is also possible that the difference in relative frontal activity at 0.5 Hz and 1 Hz compared to 3 Hz reflects the recruitment of distinct interval-timing networks. It has been suggested that regularities at faster time-scales (> 1 Hz) are processed automatically, while regularities at slower time-scales (< 1 Hz) require cognitive control (for reviews, see Lewis & Miall, 2003; Rammsayer, 2008), and recruitment of these frontal attentional networks could also explain the effect of frequency on relative frontal activity.

While the left prefrontal cortex is more dominant during lexico-semantic tasks, the right prefrontal cortex seems to be dominant during musical tasks, and during cognitively-controlled interval-timing tasks (Lewis & Miall, 2006; Koch et al., 2002). The fact that we did not observe a hemispheric difference in frontal activity may reflect shared sources across speech and music conditions, or it may reflect poor source localization due to the orientation of frontal electrical fields.

*Cortical Distortion Products*

We observed significant cortical responses at frequencies unrelated to sequences in the stimulus (Figure 3.4). These responses are not straightforwardly attributable to an envelope-tracking response. Instead, they likely reflect distortion products generated by cortical sources that process the stimulus that multiple time-scales. The fact that these responses tend to appear at 0.5 Hz intervals is consistent with this interpretation, as this pattern could emerge from a cortical source that is oscillating at both 0.5 Hz and 1 Hz (or 0.5 Hz and 3 Hz).

### *Conclusions*

We report cortical tracking of hierarchically-organized linguistic and musical structure boundaries. Since topographies of these responses largely overlap, we argue that the cortical sources recruited for the temporal processing of linguistic and musical structures may be shared. However, the fact that the linguistic stimuli used in this experiment are artificially periodic may restrict generalization to temporal processing of natural speech. Furthermore, the domain-generality of our results suggests that the cortical processes that underlie the entrainment response may play a role in the perception of any structured sound patterns.

# CONCLUSION

The experiments reported in this dissertation address three of the major issues concerning the interpretation of the cortical entrainment response to speech.

In chapter 1, we present evidence suggesting that the entrainment response follows fluctuations in speech energy within individual frequency channels of the auditory system, and that this response contributes to or reflects the differential allocation of attentional weights across these channels. While this finding does not preclude the existence of an entrainment response that follows fluctuations in energy after integration across frequency channels, it suggests that attentional enhancement of the speech target and suppression of the speech masker may be accomplished, at least in part, prior to this integration. Furthermore, it suggests that the entrainment response can be used characterize attentional weighting across frequency channels, offering the possibility of clinical application. For instance, cochlear implantation involves the insertion of an array of multiple electrodes, and the intensity of stimulation in these channels is typically determined by measuring the perceived loudness at each electrode (Thai-Van et al., 2004). However, it is not necessarily clear if these fits are optimal for speech perception, and if the electrical artifact introduced by the cochlear implant device could be sufficiently attenuated, the entrainment response measured within individual channels could be used to measure which channels are important for speech perception for individual users.

In chapter 2, we show that when acoustic confounds are controlled for, the entrainment response to the speech envelope does not depend on the linguistic content of the stimulus. This suggests that the envelope-following response is primarily driven by the acoustic properties of the speech stimulus. While it has been found that the cortical response to speech is better predicted using both acoustic-level and phoneme-level representations of the stimulus (Di Liberto et al., 2015), the extent to which linguistic content influences the envelope-tracking response has been a matter of significant debate. Furthermore, from an applied perspective, understanding the stimulus attributes that affect the envelope-following response are important, since the online measurement of linguistic information is far more difficult than the online measurement of the acoustic envelope. We also show that the strength of this envelope-following response is highly modulated by task demands, even in the absence of a competing stimulus, suggesting that attention plays a critical role in modulating the strength of the entrainment response to the speech envelope. The entrainment response to the speech envelope could potentially be used to monitor auditory attention online (Horton et al., 2014; O'Sullivan et al., 2014), and these findings might inform clinical or technological approaches that incorporate this response.

In chapter 3, we reproduce a key finding that an entrainment response can follow hierarchically-organized periodic fluctuations in linguistic information, and show that this response is not domain-specific to speech. Specifically, we show that hierarchically-organized musical stimuli can elicit a similar response pattern, suggesting that temporal processes recruited for the hierarchical processing of speech may also be recruited for the hierarchical processing of music. This is consistent with the fact that the cortical architecture for the perception and organization of musical and linguistic sounds is largely overlapping (Koelsch et al., 2002; Koelsch et al., 2004), and suggests that the temporal processes thought to contribute to the entrainment response may play a role in the perception of any structured sound patterns.

# REFERENCES

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences USA*, *98*, 13367–13372.

ANSI S3.5 (1997). Methods for the calculation of the speech intelligibility index. *American National Standards Institute*.

Baltzell, L.S., Horton, C., Shen, Y., Richards, V.M., D'Zmura, M., & Srinivasan, R. (2016). Attention selectively modulates cortical entrainment in different regions of the speech spectrum. *Brain Research*, *1644*, 203–212.

Baltzell, L. S., Srinivasan, R., & Richards, V. M. (2017). The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *Journal of Neurophysiology*, *118*, 3144–3151.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

Buchweitz, A., Mason, R.A., Tomitch, L.M.B., & Just, M.A. (2009). Brain activation for reading and listening comprehension: An fMRI study of modality effects and individual differences in language comprehension. *Psychology & Neuroscience*, *2*, 111–123.

Buzsáki, G. (2002). Theta Oscillations in the Hippocampus. *Neuron*, *33*, 325–340.

Buzsaki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*, 1926–1929.

Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., & Knight, R.T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, *313*, 1626–1628.

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain*, *125*, 1054–1096.

de Cheveigné, A., & Simon, J. Z. (2008). Denoising based on spatial filtering. *Journal of Neuroscience Methods*, *171*, 331–339.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

DéMonet, J.-F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.-L., Wise, R., … Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, *115*, 1753–1768.

Deutsch, D., Henthorn, T., Marvin, E., & Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *The Journal of the Acoustical Society of America*, *119*, 719.

Di Liberto, G.M., O'Sullivan, J. & Lalor, E. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*, 2457–2465.

Ding, N., & Simon, J.Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*, 78–89.

Ding, N., & Simon, J.Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences USA*, *109*, 11854–11859.

Ding, N., & Simon, J.Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations*. Frontiers in Human Neuroscience*, *8*, 1–7.

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

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*, 158–164.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Doelling, K.B., Arnal, L.H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85: 761–768.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*, 78–84.

Friederici, A. D. (2003). The Role of Left Inferior Frontal and Superior Temporal Cortex in Sentence Comprehension: Localizing Syntactic and Semantic Processes. *Cerebral Cortex*, *13*, 170–177.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue V. (1993). *Timit Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium.

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

Ghitza, O., Giraud, A-L, & Poeppel, D. (2013). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*, 340.

Giard, M.H., Perrin, F., Echallier, J.F., Thevenet, M., Froment, J.C., & Pernier, J. (1994). Dissociation of temporal and frontal components in the human auditory N1 wave: A scalp current density and dipole model analysis. *Electrophysiology and Clinical Neurophysiology*, *92*, 238–252.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*, 511–517.

Greenberg, S., Arai, T., & Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. *Proceedings of the fifth international conference on spoken language processing*. 74–77.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*, 465–485.

Hall III, J.W. (2007). *New Handbook of Auditory Evoked Responses*. Boston, MA: Pearson Education, Inc.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402.

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*, 3082–3093.

Horton, C., Srinivasan, R., & D'Zmura, M. (2014). Envelope responses in single-trial EEG indicate attended speaker in a "cocktail party". *Journal of Neural Engineering*, *11*, 1–22.

Howard, M.F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, *104*, 2500–2511.

Hyvarinen, A. & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483-1492.

Henry, M.J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences USA*, *109*, 20095–20100.

Hertrich, I., Dietrich, S., & Ackermann, H. (2013) Tracking the speech signal — Time-locked MEG signals during perception of ultra-fast and moderately fast speech in blind and in sighted listeners. *Brain & Language*, *124*, 9–21.

Humphries, C., Liebenthal, E., & Binder, J.R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, *50*, 1202–1211.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*, 114–116.

Kaas, J.H., Hackett, T.A., & Tramo, M.J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, *9*, 164–170.

Kerlin, J.R., Shahin, A.J., & Miller, L.M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party.” *Journal of Neuroscience*, *30*, 620–628.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 243-288). Hillsdale, NJ: Erlbaum.

Koch, G., Oliveri, M., Carlesimo, G.A., & Caltagirone, C. (2002). Selective deficit of time perception in a patient with right prefrontal cortex lesion. *Neurology*, *59*, 1658-1659.

Koelsch, S., Gunter, T. C., v. Cramon, D. Y., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach Speaks: A Cortical “Language-Network” Serves the Processing of Music. *NeuroImage*, *17*, 956–966.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, *7*, 302–307.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., & Schroeder, C.E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, *94*, 1904–1911.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., & Schroeder, C.E. (2008). Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. *Science*, *320*, 110–113.

Lakatos, P., Musacchia, G., O'Connel, M., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, *77*, 750–761.

Large, E. W., & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, *26*, 1–37.

Large, E. W. (2010). Neurodynamics of Music. In M. Riess Jones, R. R. Fay, & A. N. Popper (Eds.), *Music Perception* (Vol. 36, pp. 201–231). New York, NY: Springer New York.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–477.

Lewis, P. A., & Miall, R. C. (2003). Distinct systems for automatic and cognitively controlled time measurement: evidence from neuroimaging. *Current Opinion in Neurobiology*, *13*, 250–255.

Lewis, P. A., & Miall, R. C. (2006). A right hemispheric prefrontal system for cognitive time measurement. *Behavioural Processes*, *71*, 226–234.

Li, N., & Loizou, P.C. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *Journal of the Acoustical Society of America*, *124*, 3947–3958.

Li, F., Menon, A., & Allen, J.B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America*, *127*, 2599–2610.

Li, F., Trevino, A., Menon, A., & Allen, J.B. (2012). A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *Journal of the Acoustical Society of America*, *132*, 2663–2675.

Loizou, P.C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Jounral of the Acoustical Society of America*, *106*, 2097–2103.

Lyon, R.F., Katsiamis, A.G., & Drakakis, E.M. (2010). History and future of auditory filter models. *IEEE International Conference on Circuits and Systems* 3809–3812.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General*, *134*, 477–500.

Mesgarani, N., & Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236.

Millman, R.E., Johnson, S.R., & Prendergast, G. (2015). The Role of Phase-locking to the Temporal Envelope of Speech in Auditory Perception and Speech Intelligibility. *Journal of Cognitive Neuroscience*, *27*, 533–545.

Mondor, T.A., & Bregman, A.S. (1994). Allocating attention to frequency regions. *Perception & Psychophysics*, *56*, 268–276.

Ng, B.S.W., Schroeder, T., & Kayser, C. (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *Journal of Neuroscience*, *32*, 12268–12276.

Nie, K., Stickney, G., & Zeng, F.-G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Transactions on Biomedical Engineering*, *52*, 64–73.

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*, *31*, 10234–10240.

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *Journal of Neuroscience*, *32*, 17572–17581.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*, 713–719.

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.

G., ... & Lalor, E.C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*, 1697–1706.

Patterson, R.D. 1976. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, *59*, 640–654.

Peelle, J.E., Gross, J., & Davis, M.H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*, 1378–1387.

Phillips, D.P., Hall, S.E., & Boehnke, S.E. (2002). Central auditory onset responses, and temporal asymmetries in auditory perception. *Hearing Research*, *167*, 192–205.

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1999). Functional Specialization for Semantic and Phonological Processing in the Left Inferior Prefrontal Cortex. *NeuroImage*, *10*, 15–35.

Power, A.J., Foxe, J.J., Forde, E-J., Reilly, R.B., & Lalor, E.C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*, 1487-1503.

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

Rammsayer, T. H. (2008). Neuropharmacological approaches to human timing. In S. Grondin (Ed.), *Psychology of time* (pp. 295-320). Bingly, U.K.: Emerald Group.

Rauschecker, J.P. & Tian, B, (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences USA*, *97*, 11800–11806.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Rimmele, J.M., Zion-Golumbic, E.Z., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, *68*, 144–154.

Sanchez-Vives, M. V., & McCormick, D. A. (2000). Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nature Neuroscience*, *3*, 1027–1034.

Shahin, A., Roberts, L.E., Pantev, C., Trainor, L.J., & Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport*, *16*, 1781–1785.

Slaney, M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep* 35.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., & Davis, M.H. (2012). Predictive Top-Down Integration of Prior Knowledge during Speech Perception. *Journal of Neuroscience*, *32*, 8443–8453.

Strevens, P. 1960. Spectra of fricative noise in human speech. *Language and Speech*, *3*, 32–49.

Thai-Van, H., Truy, E., Charasse, B., Boutitie, F., Chanal, J.-M., Cochard, N., … Collet, L. (2004). Modeling the relationship between psychophysical perception and electrically evoked compound action potential threshold in young cochlear implant recipients: clinical implications for implant fitting. *Clinical Neurophysiology*, *115*, 2811–2824.

Titze, I.R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

Thompson L.W., & Thompson, V.D. (1965). Comparison of EEG changes in learning and overlearning of nonsense syllables. *Psychological Reports*, *16*, 339-344.

Tonnquist-Ulen, I. (1996). Topography of auditory evoked long-latency potentials in children with severe language impairment: The P2 and N2 components. *Ear & Hearing*, *17*, 314–326.

Vasey, M.W., & Thayer, J.F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, *24*, 479–486.

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *NeuroImage*, *172*, 766–774.

Winkler, I., Denham, S.L. & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, *13*, 532–540.

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*, 846-849.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Mechanisms Underlying Melodic Perception and Memory for Pitch. *Journal of Neuroscience*, *14*, 1906–1919.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*, 37–46.

Zion-Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*, 980–991.

Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage*, *124*, 16–23.