

ABSTRACT

Title of Document:

**MEG, PSYCHOPHYSICAL AND
COMPUTATIONAL STUDIES OF
LOUDNESS, TIMBRE, AND
AUDIOVISUAL INTEGRATION**

Julian Jenkins III, Ph.D., 2011

Directed By:

**Professor David Poeppel, Department of
Biology**

Natural scenes and ecological signals are inherently complex and understanding of their perception and processing is incomplete. For example, a speech signal contains not only information at various frequencies, but is also not static; the signal is concurrently modulated temporally. In addition, an auditory signal may be paired with additional sensory information, as in the case of audiovisual speech. In order to make sense of the signal, a human observer must process the information provided by low-level sensory systems and integrate it across sensory modalities and with cognitive information (e.g., object identification information, phonetic information). The observer must then create functional relationships between the signals encountered to form a coherent percept. The neuronal and cognitive mechanisms underlying this integration can be quantified in several ways: by taking physiological measurements, assessing behavioral output for a given task and modeling signal relationships. While ecological tokens are complex in a way that exceeds our current

understanding, progress can be made by utilizing synthetic signals that encompass specific essential features of ecological signals.

The experiments presented here cover five aspects of complex signal processing using approximations of ecological signals : (i) auditory integration of complex tones comprised of different frequencies and component power levels; (ii) audiovisual integration approximating that of human speech; (iii) behavioral measurement of signal discrimination; (iv) signal classification via simple computational analyses and (v) neuronal processing of synthesized auditory signals approximating speech tokens. To investigate neuronal processing, magnetoencephalography (MEG) is employed to assess cortical processing non-invasively. Behavioral measures are employed to evaluate observer acuity in signal discrimination and to test the limits of perceptual resolution. Computational methods are used to examine the relationships in perceptual space and physiological processing between synthetic auditory signals, using features of the signals themselves as well as biologically-motivated models of auditory representation. Together, the various methodologies and experimental paradigms advance the understanding of ecological signal analytics concerning the complex interactions in ecological signal structure.

MEG, PSYCHOPHYSICAL AND COMPUTATIONAL STUDIES OF LOUDNESS,
TIMBRE, AND AUDIOVISUAL INTEGRATION

By

Julian Jenkins III

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor David Poeppel, Chair
Professor Catherine E. Carr
Associate Professor Jonathan Z. Simon
Associate Professor William J. Higgins
Associate Professor William J. Idsardi
Associate Professor Timothy Horiuchi

© Copyright by
Julian Jenkins III
2011

Dedication

To my mother, Paula Sandra Fullerton Jenkins. And all sentient beings.

Acknowledgements

Gratitude goes to the members of the committee for assistance at various stages of my graduate school career. Thanks also goes out to the various administrative personnel (Lois Reid, Linda Dalo, Cecilia Jordan, KeCia Harper) who assisted with various queries, issues and crises. Lastly, the following people for help in writing, experimental input, analysis, collaboration, other aspects of completing the work presented and various other things (in no order of importance):

Maria Chait, Huan Luo, Mary Howard, Ariane Rhone, Phil Monahan, So-One Hwang, Brian Dillon, Mathias Scharinger, Jeff Walker, Max Ehrmann, Pedro Alcocer, Patty Shields, Norbert Hornstein, Art Popper, Sharon Staples, Cognitive Neuroscience of Language Laboratory, C-CEBH and the Departments of Biology and Linguistics

Table of Contents

Dedication.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Figures.....	vi
Introduction.....	1
M100 Responses to Two-Frequency Complex Tones.....	3
Elicitation of Audiovisual Steady-State Responses using Pseudo-Speech Signals..	5
Psychophysical Discrimination and Clustering of Ecologically Approximate Synthetic Signals.....	7
Preattentive Classification and Physiological Measurement of Ecologically Approximate Synthetic Signals using MEG.....	9
M100 Responses to Two-Frequency Complex Tones.....	12
Introduction.....	12
Materials and Methods.....	16
<i>Subjects</i>	16
<i>Threshold Testing</i>	16
<i>Perceptual Loudness Estimation</i>	16
<i>Stimuli</i>	18
<i>Delivery</i>	19
<i>Recording</i>	20
<i>Data Analysis: Peak RMS and Latency Analysis</i>	20
<i>Data Analysis: Dipole Source Estimation</i>	23
Results.....	24
Discussion.....	30
Figures.....	36
Elicitation of Audiovisual Steady-State Responses using Pseudo-Speech Signals....	45
Introduction.....	45
Materials and Methods.....	50
<i>Participants</i>	51
<i>Stimuli</i>	51
<i>Delivery</i>	54
<i>Recording and Filtering</i>	55
<i>Sensor Selection from Pre-Test</i>	55
<i>Onset Response Evaluation</i>	56
<i>SSR Analysis</i>	57
<i>Across-Participant Response Averaging</i>	57
<i>Statistical Analyses</i>	58
<i>Participant Head Location</i>	59
Results.....	59
<i>Across-Participant Power Analysis</i>	60
<i>Statistical Summary</i>	62
<i>SSR Power Comparisons</i>	64

Discussion	67
Figures	73
Psychophysical Discrimination and Clustering of Ecologically Approximate Synthetic Signals	85
Introduction	85
Part I: Psychophysical Evaluation of Signal Pairs	92
<i>Materials and Methods</i>	92
<i>Results</i>	100
<i>Discussion</i>	108
Part II: Clustering of ecologically-based synthesized signals	110
<i>Signals</i>	110
<i>Clustering Methods</i>	112
<i>Results</i>	115
<i>Discussion</i>	123
General Discussion	126
Figures	129
Preattentive Classification and Physiological Measurement of Ecologically Approximate Synthetic Signals using MEG	154
Introduction	154
Materials and Methods	162
<i>Participants</i>	162
<i>Stimuli</i>	163
<i>Experimental Procedure</i>	166
<i>Delivery</i>	168
<i>Recording</i>	168
<i>Evoked Response Analysis</i>	169
<i>Statistical Analysis of Responses</i>	170
Results	173
<i>M100 Amplitude and Latency Pretest Results</i>	173
<i>Preattentive Timbre Classification Results</i>	176
<i>/a/ - /i/ Experimental Assignment</i>	177
<i>/a/ - /u/ Experimental Assignment</i>	182
<i>Effects of Musical Training and Native Language</i>	186
Discussion	186
Figures	192
Summary	215
Bibliography	224

List of Figures

Chapter 1: M100 Responses to Two-Frequency Complex Tones

Figure 1: Stimuli Schematic

Figure 2: MEG Sensor Configuration and Contour Map

Figure 3: Single Subject Response Profile

Figure 4: Mean M100 Amplitude and Latency for Single Signals

Figure 5: Mean M100 Amplitude and Latency for Complex Signals

Figure 6: M100 RMS of RMS for Complex Signals

Figure 7: M100 RMS of RMS for Complex Signals and High Frequency
Component

Chapter 2: Elicitation of Audiovisual Steady-State Responses using Pseudo-Speech Signals

Figure 1: Construction of Audiovisual Experimental Signals

Figure 2: MEG Sensor layout, Single Participant Response Profile and Sensor
Division

Figure 3: Single-Participant Cortical Activity to Steady-State Signals
in the Temporal Domain

Figure 4: Grand Averaged Steady-State Response Profile in the
Frequency Domain

Figure 5: Density Estimates for Linear and Decibel Response Power

Figure 6: Mean Harmonic Power for each experimental condition

Figure 7: Mean Harmonic Power for each experimental condition,

Posterior Temporal and Occipital Sensors

Figure 8: Grand Average Magnetic Field Topography, Modulation Frequency

Figure 9: Grand Average Magnetic Field Topography, Second Harmonic

Chapter 3: Psychophysical Discrimination and Clustering of Ecologically Approximate Synthetic Signals

Figure 1: Spectral Envelope Structure of Experimental Materials

Figure 2: Construction of Experimental Signals

Figure 3: Proportion Correct by Vowel Class

Figure 4: Proportion Correct for /a/ -/i/ Experimental Assignment

Figure 5: Proportion Correct for /a/ -/u/ Experimental Assignment

Figure 6: Observer Discrimination Acuity Score Distributions

Figure 7: Observer Discrimination Criteria Distributions

Figure 8: Cochlear Model Representation of an Experimental Signal

Figure 9: Effect of Spatial Filters on Cochlear Representation

Figure 10: Dendrogram, Three Cluster Solution

Figure 11: Bootstrap-supported Dendrogram Clusters

Figure 12: Two-dimensional MDS Space Representations

Figure 13: Three-dimensional MDS Space Representations

Figure 14: Three-dimension MDS Space Representations, Gross Feature

Comparison

Chapter 4: Preattentive Classification and Physiological Measurement of Ecologically Approximate Synthetic Signals using MEG

Figure 1: Spectral Envelope Structure of Experimental Materials

Figure 2: Construction of Experimental Signals

Figure 3: Visualization of Timbral Metamers

Figure 4: Response Distribution of Pretest Signals

Figure 5: Mean M100 Amplitude for Pretest Signals

Figure 6: Mean M100 Latency for Pretest Signals

Figure 7: Grand Average Evoked Response Waveforms for /a/ - /i/

Experimental Assignment

Figure 8: Grand Average of Timbral Metamer Evoked Response Waveforms

for /a/ - /i/ Experimental Assignment

Figure 9: Grand Average of Synthetic Vowel Evoked Response, /a/ - /i/

Experimental Assignment

Figure 10: Grand average of the magnetic field deflections and RMS with

grand averaged topography, /a/ - /i/ experimental assignment

Figure 11: Bar plots of the mean M100 amplitude and latency with standard

errors for the /a/ - /u/ experimental assignment

Figure 12: Grand averages (RMS of RMS) of the metameric comparisons for

the /a/ - /u/ experimental assignment

Figure 13: Grand average of the magnetic field deflections and RMS with

grand averaged topography, /a/ - /u/ experimental assignment

Introduction

Processing of the sensory environment by the nervous system at all levels (periphery, ascending and descending sensory pathways, cortex) requires the parsing of complex signals, both ecological and non-ecological. For example, in the case of an auditory speech signal, the auditory pathway must analyze parameters such as the frequency content, temporal evolution and location of a signal simultaneously and often in the presence of other auditory signals that may lower the intelligibility of the signal of interest (Belin et al., 2000; Hickok and Poeppel, 2000; Hickok and Poeppel, 2007). In the case of an audiovisual speech signal, auditory and visual information must be combined across space and time to form a coherent percept (Molholm et al., 2002; Molholm et al., 2004; Amedi et al., 2005; Macaluso and Driver, 2005; Miller and D'Esposito, 2005; Molholm et al., 2006; Lalor et al., 2007; Kelly et al., 2008). The challenge for an experimenter then is not only to determine how the nervous system performs the computations required for signal parsing, but also the correct signals to employ.

Determination of the correct signals to use largely depends on the experimental questions being investigated. Simple signals (e.g. sinusoids) are useful in determining basic sensory processing (such as filter or neuronal spiking dynamics) but the results gained from such experiments may not necessarily be generalizable to a larger set of signals, especially complex ecological ones. On the other hand, signals such as audiovisual speech are too complex in a manner that exceeds the current state of the art, though they have the ability to inform about how cognition relates to signal

processing. One way to examine the interaction between basic sensory processing and cognitive aspects of signal evaluation (essentially, having one's cake and eating it too) is to employ synthetic signals that capture relevant (or salient) parameters of ecological signals, e.g., sine wave speech. These signals serve as an intermediate level between simple signals that may lack ecological relevance and the more complex, incompletely understood, ecological/environmental signals that are typically processed.

The main advantage of employing synthetic signals is that the experimenter has the ability to parametrically control specific parameters or salient dimensions as parametric control of ecological tokens is limited at best. For example, when investigating timbre (tone color) perception, using synthetic signals and then modifying specific attributes of those signals (e.g., harmonic content, envelope structure) can allow the researcher to explore the interaction between signal dimensions/parameters as well as their independence from one another (Caclin et al., 2007). Naturally, employing synthetic signals is not enough. The use of such signals must be coupled with the appropriate experimental paradigms (e.g., physiological or behavioral) and analysis techniques (e.g., evoked response component power, observer acuity) to characterize how signals of an intermediate nature inform about sensory information capture in the periphery, the complex transformations that occur in the ascending and descending sensory pathways and cognitive input used in forming a coherent percept.

The experiments presented in this thesis employ synthetic signals of an intermediate nature to explore (i) frequency and amplitude contributions of the individual components of two-frequency auditory signals and how the tone content is reflected in the major auditory evoked component; (ii) examining audiovisual integration over extended time scales using signals that contain the major features of an ecological speech signal; (iii) examine the relationship between the mathematical descriptions of auditory signals and the judgment of a human observer as they relate to timbre perception using signals that are approximations to vowels; (iv) preattentive processing of signals that are approximations to vowels. The experiments use several paradigms: magnetoencephalography (MEG) to evaluate cortical physiology, discrimination tasks to probe observer resolution and clustering and texture analyses to explore possible timbre space configurations. The first two chapters (cortical evaluation of two-frequency complexes and audiovisual integration) have already been published.

M100 Responses to Two-Frequency Complex Tones

Magnetoencephalographic (MEG) and electroencephalographic (EEG) studies have determined the response properties of the auditory evoked components. These components, namely the P50/M50, N1/N1m/M100 and P2/P2m are sensitive to such features as the bandwidth, fundamental frequency (F0), harmonic content and spacing and complexity of an auditory signal (Roberts et al., 2000; Chait et al., 2004; Shahin et al., 2005; Shahin et al., 2007; Chait et al., 2010). The most well-studied of these components is that of the N1/N1m/M100, which tends to display its peak activity at

~100 msec post-signal onset. The latency and amplitude of the M100 is modulated by parameters such as waveform shape, fundamental frequency, pitch and signal level (Poeppel et al., 1997; Roberts et al., 1998; Stufflebeam et al., 1998; Roberts et al., 2000). In the frequency domain, signals with lower detection thresholds and that the auditory system are more sensitive to (~1000 Hz) exhibit faster M100 latencies and larger RMS amplitudes than lower frequency signals (~100 Hz). Regarding signal power level, it has been demonstrated that an increase in signal level amplitude results in faster M100 latencies, especially for low-frequency signals (Stufflebeam et al., 1998).

The complex interactions between frequency content and signal level warrant further consideration due to the structure of ecological signals. Ecological signals contain components of varying frequencies and these components have varying power levels (Poeppel et al., 1997; Vihla et al., 2000; Vihla and Salmelin, 2003; Jacobsen et al., 2004a; Jacobsen et al., 2004b). The majority of M100 studies use either very simple signals (e.g., single-frequency sinusoids) or ecological signals (e.g., vowel tokens), but there are few studies investigating signals of an intermediate nature and how their processing is reflected in the M100. The first chapter describes an experiment using auditory signals consisting of two frequencies (in different critical bands and without harmonic and/or octave relationships) presented at different signal levels to examine how frequency content and power level of the components is reflected in M100 latency and amplitude (Stufflebeam et al., 1998). The experiment is designed to examine the relationship between the physical structure of a stimulus,

its perceptual attributes and the neuronal processing reflected in the M100. The deeper motivation is to understand how different features of signals are integrated across critical bands and to explore how the modulation of the M100 might be used to explore cognitive evaluation of signals.

Elicitation of Audiovisual Steady-State Responses using Pseudo-Speech Signals

The ecological speech signal usually consists of both an auditory and a visual component that must be integrated across space and time to form a coherent percept (Macaluso and Driver, 2005; Miller and D'Esposito, 2005; Lalor et al., 2007). The auditory signal typically consists of linguistic units and other extra-linguistic sounds used in communication. A reductive characterization of the auditory speech signal is that it is a broadband signal that undergoes simultaneous amplitude and frequency modulations. The visual signal generally consists of facial movements; movement of the mouth, eyes and facial muscles help to increase the intelligibility of the signal as they are paired with vocal excitation (Sumbly and Pollack, 1954). The salient aspects of the visual component of an audiovisual speech signal could then be reduced at the very least to mouth movements. The importance of having both auditory and visual information is supported by data which indicate that the pairing of a matched visual component to an auditory signal substantially decreases reaction times and increases signal-to-noise ratio, rather than vice-versa.

The benefits of having redundant information in an ecological communication signal are reflected in neuronal dynamics both in both the temporal and frequency

domains. Evoked response temporal analyses have demonstrated that redundant information affects processing in what have been considered traditionally unisensory cortical areas, exhibiting a facilitation of the latency and amplitude of evoked components (Molholm et al., 2002; Besle et al., 2004). In the frequency domain, specific frequency bands are associated with such tasks such as working memory and speech processing (Senkowski et al., 2007; Senkowski et al., 2008). Regarding the ecological speech signal, redundant information (i.e. having both an auditory and a visual component in the signal) results in an increase in power of several frequency bands (Howard and Poeppel, 2010; Luo et al., 2010). The advantage of having crossmodal information present in a signal is thus reflected in both behavioral and physiological measures of signal analysis (i.e., RT, response additivity).

A synthetic signal of an intermediate nature approximating a speech signal would need to represent the broadband carrier characteristics of the auditory component, the mouth movements of the visual component, the quasi-steady-state nature of the signal and any of the modulation inherent in ecological speech. The experiment presented in the second chapter employs three-octave pink noise (broadband carrier) amplitude modulated within the range typical syllabic rate for human speech (3.125 Hz) and paired with an ellipse modulated at the same rate to simulate mouth movements. The temporal and spectral characteristics of the evoked response are analyzed to determine how the presence of redundant information affects neuronal processing.

Psychophysical Discrimination and Clustering of Ecologically Approximate Synthetic Signals

Auditory signal processing, like sensory processing in other modalities, is perhaps fundamentally pattern recognition processing. Auditory cognition involves not only low-level decomposition along the ascending and descending auditory pathways (e.g., extracting spectral content, temporal modulations) but also higher-level inputs (e.g., phonemic information). Auditory signals are then grouped into functional categories in order for the observer to navigate the environment based on their content and structure as well as relatedness to other signals.

A full accounting of how this process is accomplished is incomplete. Previous research has established how the ascending and descending auditory pathways parse signals, but this has not led to any great insight concerning category formation (McAdams et al., 1995; Hickok and Poeppel, 2007; Okamoto et al., 2010). It is known however, that the auditory signal attribute most responsible for signal identification and hence category formation is that of timbre. Timbre is a multidimensional attribute that is integral to the identity of the source of a signal (McAdams et al., 1995; Lakatos et al., 1997; Lakatos, 2000; Meyer et al., 2006). Psychophysical data provide evidence of the salient signal dimensions that contribute to timbre perception (Plomp et al., 1967; Plomp and Steeneken, 1969; Pols et al., 1969; Caclin et al., 2007). The majority of studies focus on how spectral information contributes to timbre perception, but other dimensions (e.g., spectral spread and irregularity) depending on task demands and the signals employed also contribute.

A logical way to examine timbre perception and possible signal taxonomies would be to employ psychophysical paradigms in addition to computational analyses that create possible taxonomies based on mathematical representations of auditory signals. Several psychophysical paradigms have been successfully implemented (e.g., multidimensional scaling, forced-choice, verbal or semantic scales, Garner interference) to examine the relationships between the different dimensions in an auditory signal, while testing the limits of timbral resolution. Computational analyses are instructive in that their use can illustrate how low-level/intrinsic signal features and mathematical representations of signals, similar to the operations performed by the auditory system, can create possible functional categories and signal taxonomies (Fellowes et al., 1997; Lakatos, 2000).

The two experiments presented in the third chapter utilize a psychophysical discrimination task and simple computational analyses to probe the limits of spectral resolution regarding timbre and how signal features can combine to create functional taxonomies, respectively. The signals employed are synthetic approximations to vowel tokens in accordance with source-filter theory (Fant, 1980; Hillenbrand et al., 1995; Fant et al., 2000; Diehl, 2008); timbral differences within vowel categories are created by removing harmonics from the source waveform. The psychophysical discrimination task tests the ability of a human observer to discriminate signals based on their spectral structure (von Bismarck, 1974a,b; Luo et al., 2007), with special attention given to the discrimination (or lack thereof) of metamers – physically different yet perceptually identical signals. The clustering analyses examine how

mathematical descriptions that relate to the signal processing performed by the auditory pathway (waveform structure, power spectral density, gross structure and fine feature enhancement) give rise to possible signal taxonomies and timbre spaces. The psychophysical and computational results are then compared to see how well they agree and to what extent each paradigm is useful for investigating auditory perception.

Preattentive Classification and Physiological Measurement of Ecologically Approximate Synthetic Signals using MEG

As noted previously, the evoked response components are modulated by different aspects of the physical structure of auditory signals, with each of the components exhibiting different sensitivities to different parts of a signal's physical structure. The M50 component is sensitive to signal bandwidth and signals having larger bandwidths typically result in more robust M50 responses (Chait et al., 2004). The N1/N1m/M100 component is sensitive to features such as F0, vowel transfer function structure, signal complexity and waveform structure (Poeppel et al., 1997; Roberts et al., 1998; Stufflebeam et al., 1998; Gage and Roberts, 2000; Roberts et al., 2000; Mäkelä et al., 2004; Salajegheh et al., 2004; Tiitinen et al., 2004, 2005; Howard and Poeppel, 2009). The P2/P2m component tends to be more robust when auditory signals are more complex, e.g., the signals contain a large number of harmonics (Shahin et al., 2005; Shahin et al., 2007). Parametrically controlled synthetic signals with timbral differences are an ideal way to examine how differences in auditory signal structure modulate the latency and amplitude of evoked response components since adjustment in different attributes (e.g. harmonic content) can be used to probe

how different features are processed and how processing of these features are reflected in the evoked response component (Caclin et al., 2005; Caclin et al., 2007; Caclin et al., 2008).

Modulation of evoked response component amplitude latency does not necessarily depend on behavioral measures or some sort of participant training (Brattico et al., 2003). Rather, evoked response sensitivity to features such as spectral envelope, fundamental frequency and harmonic content are processed preattentively, even in synthetic signals sharing major features of ecological signals (Vihla et al., 2000; Vihla and Salmelin, 2003; Jacobsen et al., 2004a; Jacobsen et al., 2004b; Tiitinen et al., 2004, 2005). The latency and amplitude of the M100 simultaneously reflects fundamental frequency and spectral envelope structure and can thus be used as a measure of the differences in these dimensions between signals. More specifically, the M100 exhibits hemispheric asymmetries in the processing of signals. For example, based on neuroimaging data, it has been argued that the LH preferentially processes temporal information whereas the RH preferentially processes spectral information in speech (Poeppe and Hickok, 2004), though multiple scales are present. This has been extended to some extent for synthesized approximations presented in isolation (Tiitinen et al., 2004, 2005).

The goal of the fourth experiment is to use synthesized signals approximating vowels (identical to those used in the computational experiment) to examine to what extent spectral structure is reflected in the evoked component response. *A priori*, it is

expected that fundamental frequency, bandwidth and spectral envelope structure will be represented in both amplitude and latency for the evoked response components. The main goal of the experiment is to determine whether or not the evoked response, specifically the M100, reflects the ability to resolve subtle differences in harmonic structure, either in the component's amplitude or latency.

M100 Responses to Two-Frequency Complex Tones

Introduction

Magnetoencephalographic (MEG) studies have shown that both physical and perceptual attributes of auditory stimuli are reflected in the timing and magnetic field deflection of the major auditory-evoked component, called the N100m or M100 (Mäkinen et al. 2004; Salajegheh et al. 2004; Hari et al. 1980; Hari and Mäkelä 1988; Eulitz et al. 1995; Fujioka et al. 2003; Seither-Preisler et al. 2003; Kirveskari et al. 2006; Lütkenhöner et al. 2006). This field component, occurring ~100 msec after stimulus onset, has been shown to be sensitive to differences in the frequency, waveform shape (e.g., sinusoidal versus sawtooth), and intensity of the stimulus (see Roberts et al. 2000 for review). For example, lower frequency signals (~100 Hz), a frequency range typical of the fundamental frequency of male voices, show longer latencies and decreased field deflections than higher frequency signals (~1000 Hz), a frequency range typical for the first formant in speech (Fant 1960). Crucially, this relation holds even when frequencies are equated for loudness level. However, this interaction is not static. The latency and amplitude in response to lower frequency signals can be decreased and increased, respectively, by increasing the intensity (amplitude) of the signal (Stufflebeam et al. 1998).

It has also been argued that the processing reflected in the M100 reflects perceptual attributes of a stimulus (Roberts et al. 2000). This implies that the perceptual attributes of signals are at least partially reflected in the latency and

amplitude of the M100 component, insofar as the particular attribute can be extracted by 80 msec, the time at which the M100 is generated. To that end, it is useful to investigate the link between loudness, a perceptual attribute of an auditory stimulus with a clear and obvious contribution from the physical structure (intensity) of the stimulus (although not a one-to-one relationship), and its effect on the latency and amplitude of the M100. In short, the question is to contrast the effect on the M100 of direct physical properties of the signal and properties derived from the signal.

The principal paradigm has been to play simple single signals (e.g., sinusoids and square waves) or single speech sounds (e.g., vowels) and to evaluate the timing, amplitude, and dipole localization of the M100 (Pantev et al. 1989; Eulitz et al. 1995; Diesch et al. 1996; Diesch & Luce 1997, 2000; Lütkenhöner et al. 2001; Obleser et al. 2003, Obleser et al. 2004 a,b). On one hand, there have been numerous studies utilizing such signals that have been useful in characterizing response properties of the M100. At the other extreme, studies using ecological signals such as speech and music employ stimuli that are acoustically and spectrally complex in a way that exceeds our current understanding. However, there is not yet a rich literature on acoustic stimuli that are of an “intermediate” nature. Therefore, the linking hypotheses between the processing of simple stimuli and the processing of ecologically natural stimuli remain a bit underspecified. Signals such as speech, music, and natural and artificial sounds contain not only mixtures of frequency components but also components that contribute in unequal ways to the overall

structure of the signal, i.e., the power levels of the components may be heavily biased toward a single component or group of components.

There are two—somewhat independent—motivations to pursue a deeper understanding of the M100 and its properties. First, there is a need to understand more about auditory neuroscience in the context of human auditory processing, and specifically the integration across auditory channels and critical bands, with a view to the cortical parsing of complex signals. The features of ecological signals contain components that must be integrated across a variety of critical bands and processing channels. The M100 may be particularly well suited to elucidate relatively early cortical aspects of neuronal encoding of complex sound processing (see Chait et al. 2006, 2007 for studies on the analysis of Huggins pitch and change detection, respectively). A second motivation derives from the fact that the M100 response and its predictable variability have been used increasingly in the investigation of auditory cognition. For example, the M100 latency dependency on frequency has been used to test theories of vowel perception (Poeppel 1997; Vihla et al. 2000; Vihla & Salmelin 2002, Tiitinen et al. 2005), virtual pitch (Monahan et al. 2008), and other phenomena. Insofar as the M100 is used to test models in auditory cognitive neuroscience, studies that investigate the factors modulating it merit further attention.

To put the study into context, the mid- and long-latency auditory-evoked potentials/auditory-evoked fields have been productively used in many studies investigating auditory perception. For example, investigation of the latency of

evoked components has yielded important insights into auditory system development (Eggermont 1995), functional reorganization (Hirata et al. 1999; Brattico et al. 2003; Nikjeh et al. 2009; Okamoto et al. 2009), cochlear damage (Dietrich et al. 2001), and loudness perception and processing in both humans and nonhuman primates (Tucker et al. 2001; Tanji et al. 2010). Our study was designed to test specific hypotheses concerning loudness perception and signal component integration across critical bands. Because complex sounds contain multiple frequencies that vary in their relative power, it is important to get a more quantitative understanding of the response structure.

In this study, we used MEG, a technique well suited for the analysis of temporal information in auditory cortex (Roberts et al. 2000; Lütkenhöner & Poeppel 2010), to examine the effect on the M100 of variations in the frequency and amplitude of complex sinusoidal signals, i.e., signals composed of sinusoids of more than one frequency. Response attributes of the M100 for complex stimuli are compared with the M100 response attributes for the individual components of the complex stimuli. Intuitively speaking, the most straightforward hypothesis may be that the integrative processing (across frequency and time) required by complex signals may be associated with response prolongation, amplitude increases, and other factors reflective of the engagement of additional neural resources because of the increased complexity implicit in the perceptual analysis of structured signals. Interestingly, we observe the opposite pattern for response latency.

Materials and Methods

Subjects

Fourteen right-handed (Oldfield 1971) normal-hearing adult subjects (nine females) underwent MEG scanning. One subject was excluded from the analysis because of an insufficient signal to noise ratio for all experimental conditions. Age range was 19 to 27 yrs, mean 21.8 yrs. Subjects were compensated (\$10/hr) for their participation. Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Before the start of the experiment, informed written consent was obtained from each subject.

Threshold Testing

Threshold testing was conducted using the Hughson-Westlake paradigm (Carhart & Jerger 1959) while the subject was in the MEG shielded room. Subjects' hearing threshold was determined using 250, 500, 1000, and 2000 Hz tone pulses delivered via an Earscan 3 audiometer (Micro Audiometrics Corp., Murphy, NC). Potential subjects with 5 dB HL difference between ears or a hearing threshold 15 dB HL for any frequency were rejected.

Perceptual Loudness Estimation

Perceptual loudness was estimated by placing the subject in the MEG scanner with the experimental earphones (to simulate experimental conditions) before the start

of the experiment. Each subject was asked to discriminate a pair of signals, with both members of a pair consisting of sinusoids. Subjects had to indicate which pair of signals had the same loudness (Moore 2004). The first signal in a pair was a 1000 Hz signal presented at 60 dB above threshold. The second signal was 127, 252, or 800 Hz signal at one of the three varying levels. The levels were chosen based on previous loudness studies and standards (ISO 226:2003), which give an estimate of the SPL for the loudness of sinusoids of varying frequencies to be matched to a 1000-Hz sinusoid. For all subjects, the levels of the lower frequency signals ranged from 50 to 75 dB SPL. The three levels were as follows: (a) 5 dB below the SPL most likely to provide a match to the 1000 Hz signal, (b) at the sound pressure level most likely to match that of the 1000 Hz signal, and (c) 5 dB above the sound pressure level most likely to match the 1000 Hz signal. All signals were of 1 sec duration and had a 7 msec \cos^2 onset and offset. Subjects were first asked to evaluate the loudness of the signal pairs independently as many times as needed and then to respond via button press when the two signals had the same perceptual loudness. If the signals were not matched in loudness then the level of the volume of the lower frequency signal was adjusted by the experimenter, and the subject was asked to repeat the task until the perceptual loudness was matched. The signals with the same perceptual loudness were carried over and used as the 0 dB signals for the experiment (see below).

Stimuli

Signals for the experiment were generated with MATLAB (v7, The Mathworks, Natick, MA). Four sinusoidal signals of the frequencies 127, 252, 800, and 1000 Hz were sampled at 44.1 kHz with 16-bit resolution. Signals were generated using the sine function not the cosine function. Duration was 400 msec with a 7 msec \cos^2 onset/offset ramp. Complex signals consisted of combinations of 127 Hz with 1000 Hz and 252 Hz with 800 Hz. Individual frequencies and their combinations were selected to be in non-overlapping critical bands.

The two lower frequencies were presented at five different signal levels (with respect to the 800 and 1000 Hz signals 60 dB above the subject's threshold). A reference value of 1000 Hz was used because of the human auditory system's sensitivity to frequencies between 1000 and 2000 Hz, as well as the well known M100/N100m and associated parameters for this frequency. The relative values used were -20, -10, 0, +10, and +20 dB. The two high-frequency signals, the 800 and 1000 Hz signals, were only presented at one loudness level (0 dB). The two complex signals were presented in mixtures of amplitudes for a total of 22 signals (12 single signals and 10 complex signals; see Fig. 1). The energy of the complex signals and simple signals was not identical because it was important to keep all signal components that constituted the complex signals identical to their simple signal counterparts to accurately gauge the effect on M100 latency and peak root mean square (RMS). Each stimulus was presented 100 times, pseudorandomly interleaved. Complex signals were more akin to "chords" rather than being mathematically

additive; as such, special care was taken to make sure that both ears received the same signal being applied. The experimental materials were passively listened to; no response was required from the subjects. To maintain the vigilance of the subjects, a distracter task was incorporated into the experiment. Approximately Gaussian white noise (400 msec duration) was used as the target during the experiment and was pseudorandomly presented with the signals (17% of total). Subjects had to press a button in response to the noise target; these trials were excluded from analysis.

Delivery

All experimental stimuli were presented using a Dell Optiplex computer with a SoundMAX Integrated HD sound card (Analog Devices, Norwood, MA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Stimuli were delivered to the subjects binaurally via Eartone ER3A transducers and nonmagnetic air tube delivery (Etymotic, Oak Brook, IL). The inter-stimulus interval varied randomly in the time interval between 300 and 700 msec. The decision to use a relatively short interstimulus interval (ISI) was based on practical considerations; the use of such an ISI ensured that subjects would not be in the scanner more than 1 hour, after which considerable discomfort may occur. Previous studies have investigated the impact of ISI on the peak and latency of the N1/N1m (Hari et al. 1982) with shorter ISIs exhibiting longer latencies. However, because it takes 10 to 12 secs for the response to the previous stimulus to disappear completely, extending the ISI may be artificially inflating the M100 response. Subjects were instructed to stay as still as possible while positioned inside of the scanner; no

mechanical measures were taken (such as a foam neck support) to maintain that position. However, the amount of deviation between the start of the experiment and its conclusion was quantified by taking measurements of the subjects' head positions using marker coils.

Recording

Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was DC-100 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. Data were noise reduced using time-shifted principal component analysis (de Cheveigné & Simon 2007), trials averaged offline (artifact rejection ± 2.5 pT), bandpass filtered between 0.03 and 20 Hz, and baseline corrected over the 100-msec prestimulus interval. The relatively low upper frequency cutoff value is not unusual; auditory-evoked field studies generally have filter cutoff values of 20 to 30 Hz (see Tiitinen et al. 1993; Eulitz et al. 1995). In addition, time-shifted principal component analysis does not remove eye-blink artifacts, but trials with excessive frontal field deflections were excluded from the analysis.

Data Analysis: Peak RMS and Latency Analysis

Selection of maximally responsive auditory channels was performed in a pretest, using 250 and 1000 Hz sinusoidal signals of 400 msec duration presented at 65 dB SPL. On average, the 1000 Hz signal resulted in a stronger auditory response. Five channels from source and sink from each hemisphere (i.e., 20 channels total)

with the maximum measured magnetic field deflection to the 1000 Hz signal were used for subsequent analysis. M100 peak RMS amplitude and latency (search window: 90 to 210 msec after stimulus onset) for each signal for each hemisphere were determined using the localizer channels. M100 peak RMS and latency data values for each hemisphere for each condition were averaged across the subjects (see Figs. 2 and 3 for visualization). The signal evaluation window ranged from 100 msec pretrigger to 600 msec posttrigger. M100 peak RMS time values for all data channels were collected and averaged across subjects for each stimulus and were plotted topographically to confirm the M100 response. In addition, when a subject's data did not show an auditory cortex magnetic field topography for a given condition, after averaging and filtering, that data (peak RMS and latency) was excluded from further analysis (Luo & Poeppel 2007). A total of 31 such exclusions were made.

Across-subject responses for all subjects showing a quantifiable M100 were characterized by collecting the individual RMS vectors into a matrix (according to condition) and then calculating the RMS of that matrix. The individual RMS values were calculated from the sensors selected from the pretest data for all time points of the observed response. The peak RMS value and latency for the across subject response was quantified in the same time window as the individual subject data. The motivation for using the RMS of the RMS (see Fig. 7) is as follows: (a) MEG is a technique well suited for within-subject analyses and (b) because the sensor positions are not in the same place for all subjects, unlike EEG where there are consistent points of reference, grand averages are not as straightforward. Although some

researchers do present grand averaged data, the interpretation of the results still requires RMS. Because of the inherent difficulty of presenting grand averages of channels in MEG data because of the fact that channels are not in canonical positions, we use the RMS of RMS measure to present grand averages over a derived measure (RMS) that provides an index of regional activation. This has proven a valuable and valid measure in previous studies (Chait et al. 2006).

To assess the significance of the effect of auditory signal manipulation on the M100 latency and peak RMS, the values were collected and analyzed using repeated measures analysis of variance (ANOVA) (SPSS 16.0, SPSS Inc., Chicago, IL). Missing M100 values (see above) were replaced with the across-subject series mean for the particular signal type. A full factorial design was employed with amplitude and latency as the dependent measures and hemisphere (right hemisphere (RH) versus left hemisphere (LH)), signal type (simple versus complex), and loudness level as the factors. Three additional planned comparisons were performed using paired *t*-tests. The first compared simple low-frequency signals with complex signals (e.g., 127 Hz 20 dB to 127 Hz 20 dB/1000 Hz) for the M100 latency in the same hemisphere, the second compared the same kind of signal across hemispheres (e.g., 127 Hz 0 dB RH versus 127 Hz 0 dB LH), and the third analysis compared high-frequency signals with complex signals (e.g., 800 –252 Hz 10 dB/800 Hz) within the same hemisphere. All statistical effects (repeated measures ANOVA and paired *t*-tests) are reported as significant at $p < 0.05$.

Data Analysis: Dipole Source Estimation

Because (a) the M100 response should not solely be defined by its peak amplitude and latency and (b) the evidence that stimulus frequency may influence the sources of the M100, we conducted an analysis of the estimated position of the M100 sources. Dipole sources were estimated based on data from five subjects who exhibited quantifiable M100 responses for a majority of conditions and from whom we had robust digitized head-shape data. We did not have access to structural magnetic resonance images from our subjects and are consequently unable to perform individual anatomically constrained localizations. Single equivalent current dipole estimates with a goodness-of-fit less than 80% were excluded from subsequent statistical analyses. A simple spherical head model was used to determine the source of the M100 response (x , y , z axes) as well as the dipole angles (Φ and θ) using 40 channels per hemisphere (20 source and 20 sink) with the greatest magnetic field deflection for a single moving dipole. Given their differences in spectral composition, we performed a single equivalent current dipole analysis to assess whether the different signals were associated with measurably different source localizations. The dipole modeling, performed on individual subjects head shapes using a spherical model, implemented in the MEG160 platform, yielded x , y , and z coordinates of the dipole, angles Φ and θ , as well as dipole moment in nAm. Statistical significance of the values of x , y , z , Φ , and θ were assessed using a mixed effects ANOVA and Wilcoxon signed-rank tests in R using the “languageR” statistical package (R Foundation for Statistical Computing, v 2.8.1; Baayen 2008).

Wilcoxon tests were performed on the values of Φ and θ both across and within subjects to assess any potential differences in the source orientation of the M100 response. The nonparametric tests were performed on single-frequency data values versus complex-frequency data values, e.g., 127 Hz versus 127/1000 Hz and 1000 Hz versus 127/1000 Hz. We employed the more conservative nonparametric tests because we could only use the data for 5 of 13 subjects for dipole localization. Because of such a small sample size out of the original subject pool, we thought it was more correct to assume that the data were not parametric.

Results

The stimuli typically elicited a robust M100 response, although the quality of the M100 varied as a function of signal type. Figure 2 illustrates a typical evoked field distribution recorded from 157 channels for one subject presented with one complex signal. The displayed (averaged and filtered) data show the prominent magnetic field deflection in both RH and LH channels. The peak (and trough) responses demonstrate the canonical distribution associated with the sensor configuration used in the current experiment (Fig. 2a). The spatial extent of the response, and the putative temporal lobe origin, is well captured by the absence of evoked activity along the midline channels. Calculating the topographic distribution of the response at the peak (Fig. 2b) provides evidence that the M100 recorded here yields the standard spatial response pattern observed for most auditory signals.

The butterfly plot depicted in Figure 3a shows the response to a 252 Hz signal at the +20 dB presentation level for one subject. Both the individual channel responses (black) and the aggregate response across channels (RMS, red) point to the large peak slightly after 100 msec after stimulus onset. The quantitative analysis across experimental conditions that we perform here is based on a more selective group of channels. Figure 3b illustrates the dependent variable. The left and right panels show channels from the LH and RH, respectively. As described above, we selected five channels from the source and sink of each hemisphere (black lines) and calculated the RMS response (red) across the 10 selected channels in each hemisphere. The subsequent analyses here focus on the RMS peak derived from this visualization of the data.

Figure 3c shows the RMS for a 252 Hz signal presented at the five different loudness levels. The RMS curves show clearly the relationship between the peak response and loudness; in particular, for this subject and these data, one observes a clear grouping, with the two lowest loudness values being associated with significantly lower and later M100 peaks, replicating previous findings on loudness effects on the M100 (Stufflebeam et al. 1998).

Figure 4 summarizes the mean latency and peak M100 values across subjects for the 127, 252, 800, and 1000 Hz signals at the equal loudness (0 dB) level. The 800 and 1000 Hz signals demonstrate decreased M100 latencies and increased peak M100 values relative to the 127 and 252 Hz signals. Although the peak RMS values

across hemispheres show no significant difference (Fig. 4a), the RH shows significantly decreased latencies relative to the LH for all frequencies, replicating previous data.

Figure 5 summarizes the mean peak RMS and latencies for simple low frequency and complex signals. Relative to simple signals, the complex signals show a significant increase in the peak RMS at all loudness levels for both sets of complex signals (127/1000 Hz and 252/800 Hz; Figs. 5c,d). Complex signals also show decreased latencies relative to simple signals; however, for the loudest simple signals and complex signals, the latency values are approximately the same (Figs. 5a,b). This decrease in M100 latency is contrary to the intuitive prediction that complexity would yield slower latency values because of increased cortical processing as a result of allocating more neuronal resources to parse the incoming signal. However, it has been shown in macaque auditory cortex that signal complexity decreases the latency of physiological responses (Lakatos et al. 2005). In addition, the time range of the M100 latency across loudness levels is not as great for the complex signals as opposed to the simple signals (~130 to 180 msec for simple signals, ~120 to 140 msec for complex signals). As with simple signals, the latency of the M100 is faster in the RH than in the LH. There were two significant main effects regarding M100 peak latency for the ANOVA that replicate findings from previous studies, namely those of hemisphere ($F(1,12) = 28.397$) and loudness level ($F(2.125,25.504) = 20.246$). The data show that the RH yields faster M100 peak latency values than the

LH and that as the level of the signal increases, the M100 peak latency decreases (see Figs. 5a,b).

The core result of the ANOVA is the finding that complex signals yield faster M100 peak latencies than low-frequency simple signals ($F(3,36) = 23.730$). Across all loudness levels, the complex signals generated faster mean M100 values, although not all the mean latency values were significant (see below). There were two interaction effects: a two-way interaction effect of signal type loudness level ($F(12,144) = 4.726$) and a three-way effect of hemisphere signal type loudness level ($F(6.282,75.387) = 3.088$). The complex signals generated faster M100 latencies at low signal levels (-20 to 0 dB) but not for the higher signal levels. The three-way interaction indicates that the two-way interaction was confined to the RH.

In addition, the planned comparisons involving paired t -tests on the M100 peak latency for simple low-frequency signals versus complex signals (within hemisphere) indicated that on average, the simple signals, except at the highest loudness levels, generated slower M100 peak latency values relative to the complex signals. All values for the paired comparisons and their significance were greater than $t(12) = 3.46$ and $r = 0.707$ for the RH and $t(12) = 2.63$ and $r = 0.604$ for the LH.

Figure 6 summarizes the across subject group response in terms of the peak and latency of the RMS of individual subject RMS vectors. We observe that the pattern for the maximum RMS value and its latency hold for both the simple and

complex tones, for all combinations explored. Crucially, the loudest and fastest latency simple signals (both low and high frequencies) peak at ~130 msec, except for the 800 and 1000 Hz signals in the RH, where the M100 response peaks at ~120 msec. As with the mean peak latency ANOVA, there were several main significant results regarding the mean peak RMS value, those of signal type ($F(3,36) = 35.086$), loudness level ($F(4,48) = 34.644$), and the two-way interaction of signal type loudness ($F(12,144) = 6.729$). No paired comparisons between simple and complex signals across hemispheres were performed on the peak RMS value because there was significant overlap upon graphical inspection and that there was no significant difference between hemispheres resulting from the ANOVA analysis (see Figs. 5c,d).

The experiment shows that as signal level increased, the M100 peak RMS increased and the latency decreased. For 127 and 252 Hz single signals, the range of the M100 latency was about the same (~130 to 180 msec), with no obvious lateralization. The effect size for the M100 amplitude is the same for both the LH and RH for the 127 and 252 Hz signals. For both single and complex signals, for both hemispheres, the M100 latency seems to converge on the same approximate value (~125 msec), indicating that the latency reaches a possible point of saturation.

Mean M100 latency (~130-180 msec) across subjects for 127 and 252 Hz single signals at all loudness levels, for both hemispheres, was approximately the same. As the signal amplitude increased, the M100 latency decreased and peak RMS value increased.

Additional ANOVAs were performed on the data without excluding any data. For the ANOVA performed on the latency of the M100, significant effects were found for hemisphere ($F(1,12) = 16.754$), signal type ($F(3,36) = 10.427$), and loudness level ($F(1.951,23.407) = 5.527$). No significant two-way and three-way effects were found. For the ANOVA performed on the peak RMS of the M100, significant effects were found for signal type ($F(3,36) = 48.221$), loudness level ($F(2.194,26.331) = 33.621$), and signal type loudness level ($F(12,144) = 10.651$). No other significant interactions were found. Analyses with and without replacement both found significant effects of hemisphere and loudness level for M100 latency and signal type, loudness level, and type level for the RMS.

In no spatial dimension was there a systematic effect of stimulus on the localization. This is consistent with (i) the resolution of the method and model employed, and (ii) the fact that even expected spatial differences (e.g., between 250 and 1000 Hz) are not at all reliable (see Lütkenhöner 2003 for extensive data and discussion). The mixed effects ANOVA performed on the dipole source data found only a significant interaction of hemisphere ($F(1,152) = 5112$), not surprising because of the independence of hemispheric responses. No other significant responses were found. The across-subject and within-subject Wilcoxon signed-rank tests did not find any statistical differences in the orientation of the M100 source.

Discussion

The major auditory-evoked component, the M100/N100m, exhibits variation in both the magnetic field amplitude and the peak latency as a function of signal frequency and level. In this study, we expand on the dominant paradigm to investigate to what extent the relative level of components and complexity of a signal affect the magnetic field deflection and peak latency of the M100. The signals for the current experiment are an important intermediate level of spectral complexity between simple sinusoids and more ecologically valid signals such as vowels or musical tones. Typical MEG studies focusing on the M100 component use either simple sinusoidal signals or more complex speech tokens (e.g., synthetic vowels with three sinusoidal or approximately sinusoidal components) but do not examine how the individual simple components of a complex signal contribute to the overall M100 response of the complex signal. The data for the simple signals replicate previous findings, namely that as the signal level increases, the field deflection increases and the peak latency decreases and that the RH exhibits faster latencies than the LH (Stufflebeam et al. 1998; Lütkenhöner & Klein 2007; Howard & Poeppel 2009). The analysis was performed in sensor space, not source space, to stay as close as possible to the recorded data without making source configuration assumptions. However, the sources of the auditory M100 are well understood (Lütkenhöner & Steinstrater 1998), and hypotheses about localization are not at issue in this experiment. Although we did not obtain anatomic magnetic resonance images for subjects and can therefore not perform high-resolution anatomically constrained source localization on individual subjects, the contour maps for both simple and complex signals are indistinguishable

from the patterns in the prior literature investigating the cortical sources underlying the M100. The cortical sources lie in the superior temporal lobe and are likely to include multiple distinct generators. The data we report here are highly likely to be associated with the same cortical generators. We were able to verify that there were no signal-driven source differences when using single equivalent current dipole modeling with a spherical head model derived from individual subject head shapes.

The core finding of this study is that complex signals demonstrate increased amplitude but decreased peak latency values. The finding that the magnetic field amplitude increases with complexity is not surprising: an increase in the amount of signal across channels would drive more neurons, and an increase in activity seems a reasonable hypothesis. What was unexpected is that complexity does not yield longer, slower latencies in cortical populations as a consequence of the integration across critical bands. This could be for several reasons. First, it could be that the signals are parsed and integrated very early in the ascending auditory pathway and the results of this early analysis are reflected in the response of cortical neurons. The signals used contained components well separated in frequency (in different critical bands, no harmonic or octave relationships), and from a physiological and perceptual standpoint, the signals “appear” to be two separate sinusoids played simultaneously. Such a separation could theoretically aid in the analysis of the signals starting from the cochlea and proceeding to cortical neurons.

The data also suggest that certain frequencies may have a privileged analysis by the auditory system (i.e., the excitation thresholds for these frequencies are relatively low). This is evidenced by the finding that even at perceptually matched loudness levels, the 800 and 1000 Hz signals elicit greater response amplitudes and shorter M100 peak latencies than the 127 and 252 Hz signals. The preferential treatment (lower thresholds) given to these frequencies may contribute to the decreased M100 peak latencies of the complex signals used in the experiment. This view is supported by the across-subject data presented in Figure 7. Although the data presented in Figure 7 suggests that the latency of the complex signals is wholly determined by the higher frequency component, the results of the ANOVAs for latency found statistically significant effects for signal type and loudness level, a finding that cannot be completely discounted. Even when both the power and the perception of the complex signals was dominated by the lower frequency component, the latency of the complex signal is closer to that of the higher frequency component. This creates an interesting dissociation between perception and the physiological response (see below).

Another possible explanation for the data is related to the integration window of the M100 and the amount of sound energy located in that window. The physical structure of the signal and any modulations to the signal within the first 30 to 40 msec after onset can have profound effects on the latency, amplitude, and shape of the M100 (Gage et al. 2000). The complex signals have relatively more energy in

this time period than the simple signals. The latency and amplitude facilitation may be analogous to that seen with pulse data (Sams et al. 1993).

An alternative scenario is that because the complex signals used here more closely resemble ecologically valid signals (e.g., vowels), cortical neurons are more “tuned” to them (Jacobsen et al. 2004), and this facilitates more robust and faster processing. It should be noted that the signals used contained only two frequency components as opposed to the three normally used with vowel tokens. Perceptually, this is a significant difference. Once three or more sinusoidal or approximately sinusoidal components are used, different perceptual effects can occur, based on the frequency grouping of the components as well as presentation of the signals (e.g., modulation and harmonic structure). As a kind of “halfway point” between simple sinusoids and the complex signals encountered normally in the environment, the signals shed light on the interactive effects of the processing reflected in the M100 response. The signals used also give evidence of the interactive effects associated with the processing (hemisphere, signal type, level, etc.), which would be harder to disentangle if the same method in this experiment was applied to more complex signals. However, this processing has a natural limit as evidenced by the apparent saturation of peak latency values (see Figs. 5a,b) whereby the values for both simple and complex signals converge onto the same approximate value.

Previous data suggest that different signal frequencies have different M100 source orientations, with low-frequency signals having the steepest angles in the

sagittal plane and high frequencies the shallowest angles (Tiitinen et al. 1993). The results of the mixed-effects ANOVA and the Wilcoxon tests on the phase angles of the responses did not replicate this finding. This could be due to an inability to resolve these components because of the spatial limitations of MEG (Lütkenhöner & Steinstrater 1998; Lütkenhöner 2003). However, previous studies have found that signal level and context may affect the localization of the source (Okamoto et al. 2009; Tanji et al. 2010). The primary purpose of this experiment, however, was not a focus on source localization but rather quantifying direct, surface measurements. Furthermore, even though we were not concerned with differences in hemispheric processing for the M100 for this particular study, it is vital to consider the data from each hemisphere separately for several reasons. First, the M100 literature shows that the RH exhibits faster response latencies than the LH as well as a greater power at the peak. Second, there is evidence that the hemispheres evaluate different aspects of the auditory signal. Finally, the cytoarchitecture and the cortical folding between hemispheres differ. Thus, it would be safer to evaluate each hemisphere as a separate entity for statistical purposes.

One finding of particular interest is that of the dissociation of perception from the physiological response that we are quantifying. It was previously proposed that the M100 reflects perceptual attributes of the signal rather directly (Stufflebeam et al. 1998). In this experiment, the lower frequency components at their loudest levels (+10 and +20dB) completely dominated the perception of the complex signals. However, the latency and the peak RMS values were not closer to the values for the

simple signals at those levels presented alone. The data suggest that the loudness of a signal is not entirely reflected in the M100 in the case of complex signals or at least that there is another mechanism that assists in assessing the loudness of signal and its individual components. At the very least, loudness cannot be “read out”—from a neural coding point of view—by decoding properties of the M100. This is consistent, of course, with stable loudness estimates developing over a period of about 200 msec (cf. Moore 2004).

Further studies into this topic should focus on the different grouping of signals, e.g., complex signals composed of 400 and 720 Hz components, or even components where frequencies are in the same critical band (Soeta & Nakagawa 2006). The facilitation of a complex signal and the proximity of its latency to that of a particular component may depend on the constituent frequencies, the position of those frequencies relative to hearing bandwidth, and their distance (logarithmic, linear, and log-linear) from each other.

Figures

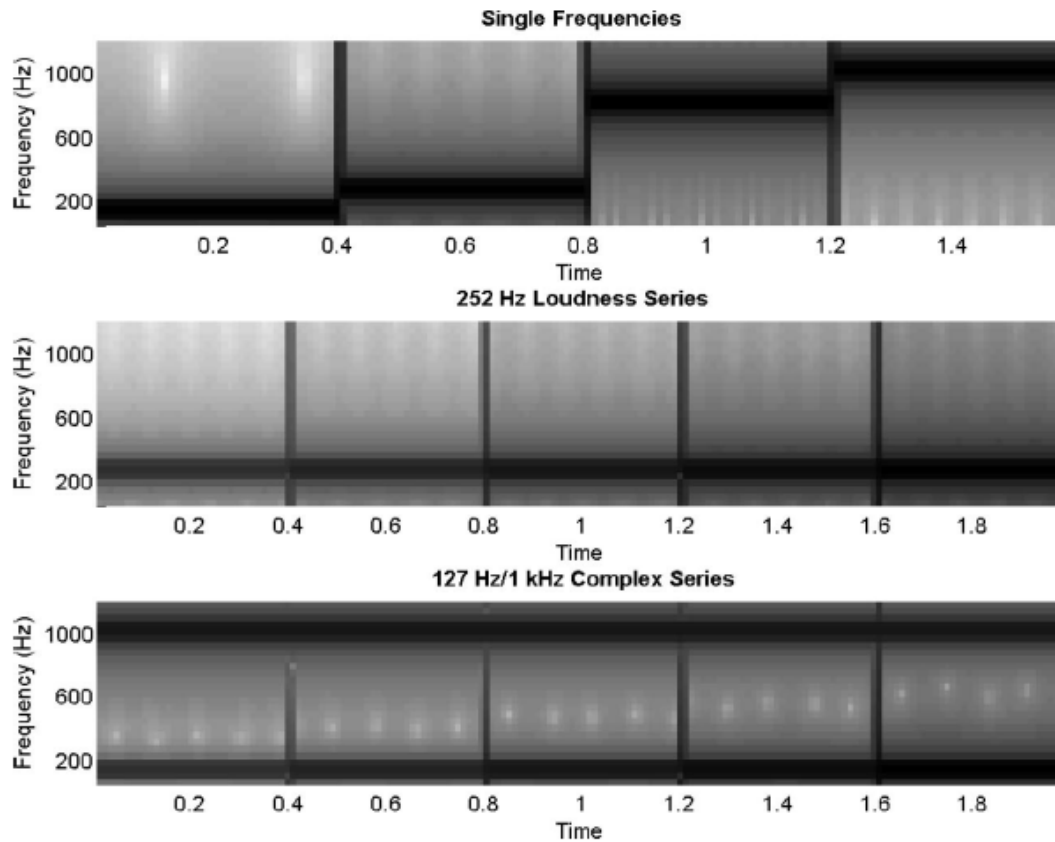


Figure 1. Schematic spectrograms of stimuli. Top panel, spectrograms of single sinusoids (left to right: 127, 252, 800, and 1000 Hz) at 0 dB loudness level. Middle panel, loudness series for 252 Hz single sinusoid from -20 dB at the far left to +20 dB at the far right. Bottom panel, loudness series for 127/1000 Hz complex sinusoid. Panels show full range of lower frequency component of stimuli from -20 to +20 dB levels.

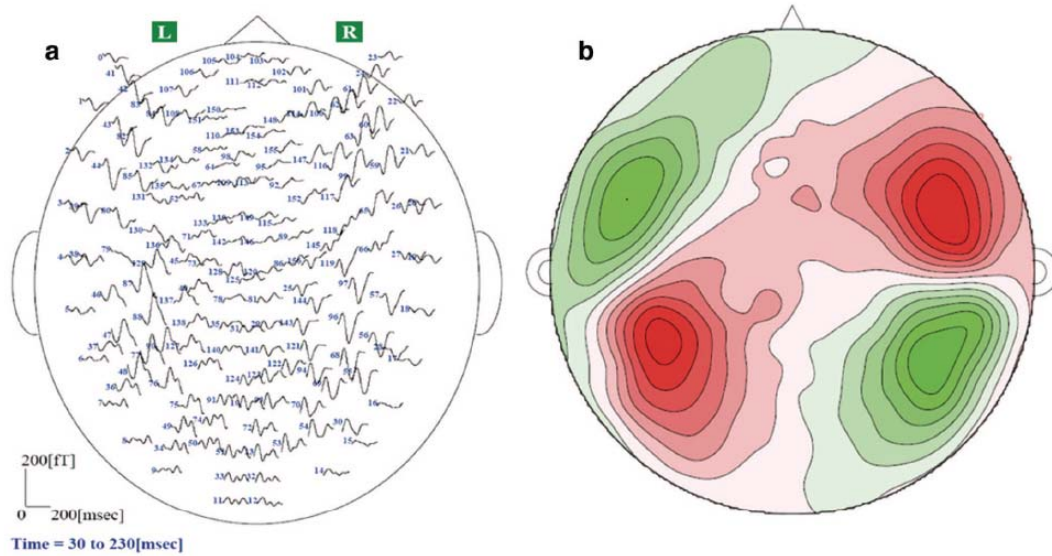


Figure 2a. Sensor configuration of whole-head biomagnetometer. The response of 157 data channels for one subject's response to a 252/800 Hz complex signal is shown. The peaks and troughs in the evoked magnetic field illustrate the typical distribution of an auditory-evoked response measured by axial gradiometers. The latency of M100 peak is 122 msec. As expected, channels along the midline, i.e., not overlaying temporal lobe cortical areas, do not reflect any interpretable evoked activity elicited by an auditory signal. Figure 2b. Contour map of the response peak of the sensor data shown in (a). Magnetic field source is in red and sink is in green, depicting the canonical left vs. right dipolar configurations.

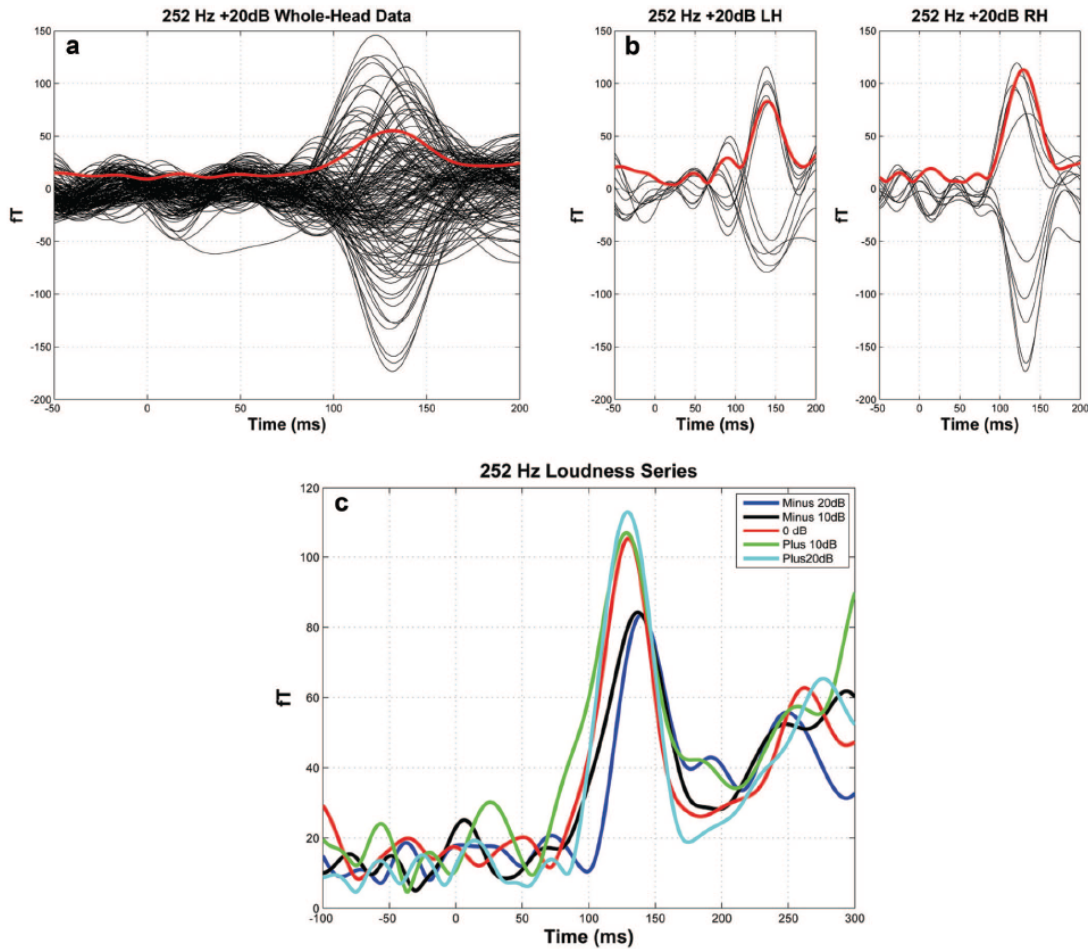


Figure 3a. Data from single subject for 157 channels in response to 252 Hz single signal at +20 dB presentation level. Black lines are magnetic field deflections from each channel; red line shows the root mean square (RMS) across all data channels. The data shown are averaged (88 trials) and bandpass filtered (0.03 to 20 Hz). This “global” view of the response (across hemispheres) illustrates the large and robust response profile elicited by the stimuli and provides a sense of the scale of the response under consideration in terms of timing and amplitude. Figure 3b. Response broken down by hemisphere. For each hemisphere, the 10 best channels (maximum field deflection), 5 from the source and 5 from the sink, are selected and their RMS calculated. Data from single subject shown in (a) for 252 Hz +20 dB single signal.

Left/right panels are the left/right hemispheres. Figure 3c. Data from single subject in (a) and (b) for 252 Hz single signals at all loudness levels (-20, -10, 0, +10, and +20 dB) from the 10 right hemisphere channels with the maximum magnetic field deflection. Displayed are the RMS values with peaks corresponding to M100 for each signal. Signals with lower intensity levels (-20 and -10 dB) show delayed latencies and lower peak M100 amplitudes relative to other signals in the series. Latency values: 140 (-20 dB), 137 (-10 dB), 129 (0 dB), 128 (+10 dB), and 128 (+20 dB) msec.

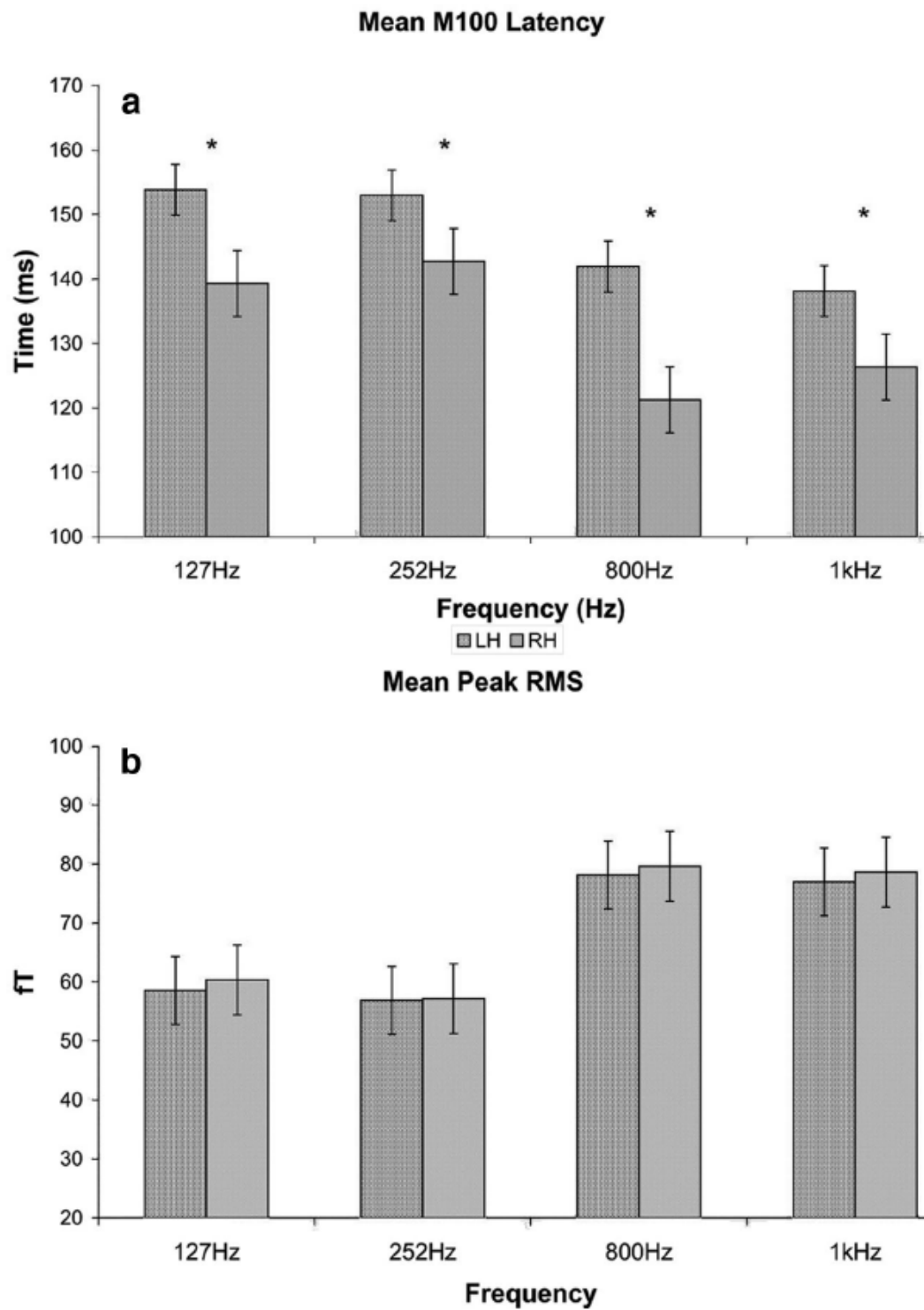


Figure 4a. Mean M100 latency across subjects for single signals at matched loudness levels (matched sensation level). Latencies for the right hemisphere are faster than

those for the left hemisphere. The latency pattern replicates a number of previous reports that show a striking latency-frequency dependence independent of sensation level (see Roberts et al. 2000 for review). Figure 4b. Mean peak root mean square (RMS) values for M100 across subjects for single signals at matched loudness levels. Peak RMS values are approximately the same in each hemisphere (i.e., effect size is about the same).

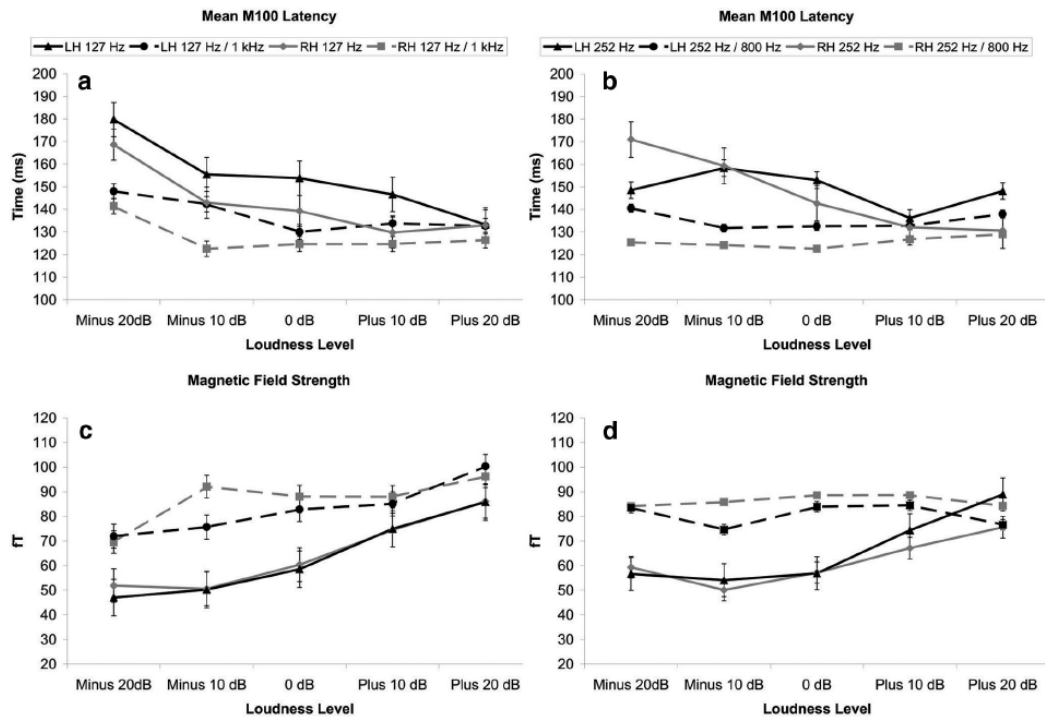


Figure. 5a. 127 Hz single signal and 127/1000 Hz complex signal M100 latencies across subjects, for both hemispheres, plotted separately. Simple signals are indicated by the solid bold lines and complex signals by the dashed lines, the RH values are indicated by gray color and the LH by black color. Data are the subject means at each loudness level for single signals and mixture levels for complex signals. The means

and SEs of the means presented exclude subject data that did not generate the M100 response. As loudness level/mixture level (signal intensity and/or power) increases (see Materials and Methods section for details), the M100 latency decreases. Data for single signals (solid lines) replicate findings from Stufflebeam et al. 1998 (latency decreases with increasing signal intensity and/or loudness). Complex signals (dashed lines) also show decreased latency relative to single signals. This pattern is contrary to the possible prediction that complex signals yield longer latencies. Figure 5b. The same data for the 252 Hz single signals and 252/800 Hz complex signals. Figure 5c. Mean peak RMS values across subjects for 127 Hz single signals (solid lines) and 127/1000 Hz complex signals (dashed lines). Relative to single signals, complex signals generate an increase in the M100 amplitude across all mixture levels. The effect size is the same in both the left hemisphere and right hemisphere. Figure 5d. Analogous to the data in (c) for the 252 Hz single signals and 252/800 Hz complex signals.

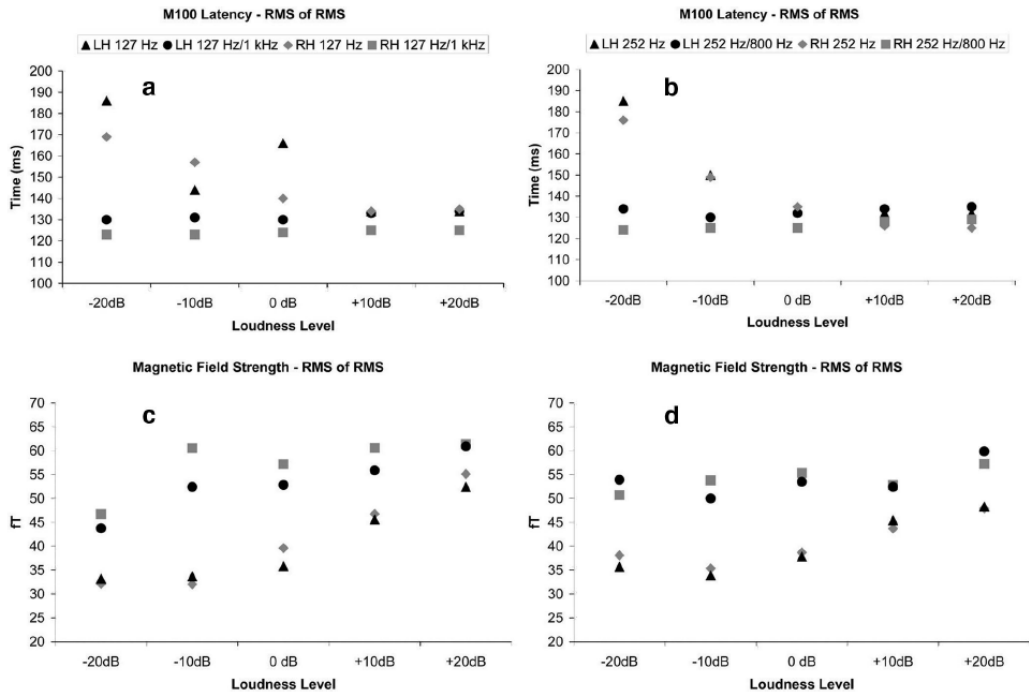


Figure 6a. 127 Hz single signal and 127/1000 Hz complex signal M100 latencies across subjects, for both hemispheres, plotted separately. Conventions used are the same as in Figure 5. Data are the RMS of subject RMS vectors collected for subjects demonstrating a quantifiable M100 for a specific condition. Across-subject RMS responses show the same response pattern as the mean values for the peak and latency of the M100. Figure 6b–d. The same data as their counterparts in Figure 5, except that mean values are replaced with across-subject RMS values.

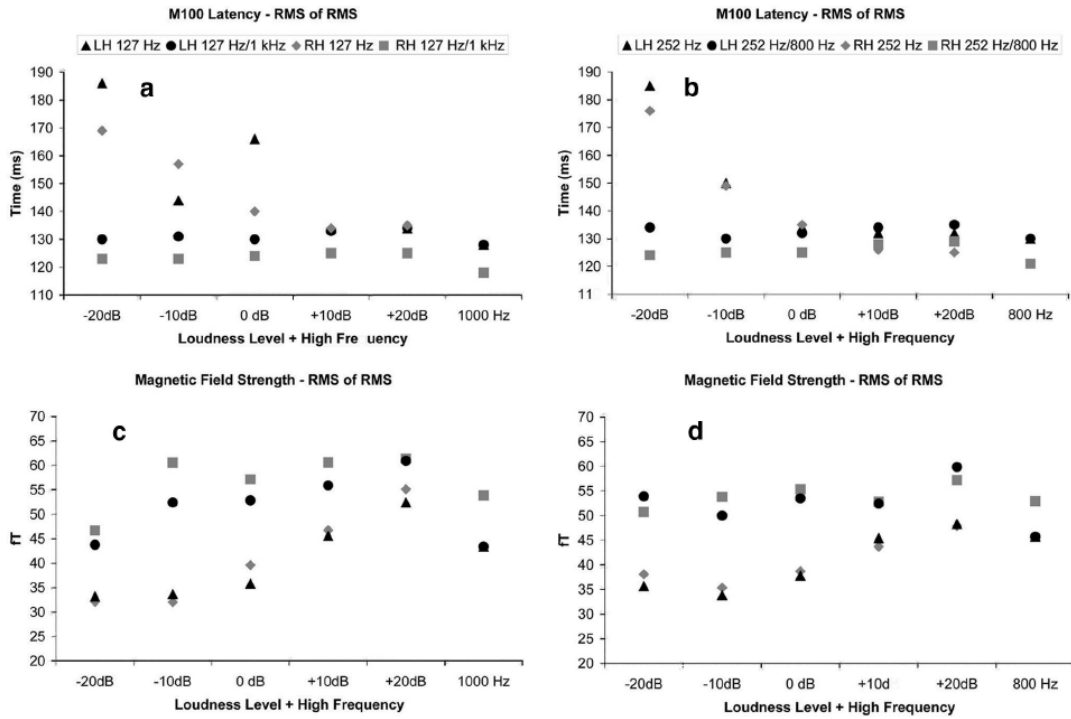


Figure 7a. 127/1000 Hz complex signal and 1000 Hz single signal M100 latency values for both hemispheres. Conventions used are the same as in Figure 5. Latency values are the RMS of subject RMS vectors collected for subjects demonstrating a quantifiable M100 for a specific condition. The latencies of the complex signals are closest to that of the 1000 Hz signal. Figure 7b. The same data for the 252/800 Hz complex signals and 800 Hz single signal. Figure 7c., Peak RMS values of across-subject RMS vectors for 127/1000 Hz and 1000 Hz single signals for both hemispheres. Figure 7d. The same data for the 252/800 Hz complex signals and 800 Hz single signal.

Elicitation of Audiovisual Steady-State Responses using Pseudo-Speech Signals

Introduction

The majority of naturalistic sensory experiences require the observer not only to segregate information into separate objects or streams but also to integrate related information into a coherent percept across sensory modalities - as well as across space and time (Amedi et al. 2005; Kelly et al. 2008; Lalor et al. 2007; Macaluso and Driver 2005; Miller and D'Esposito 2005; Molholm et al. 2002, 2004, 2007; Murray et al. 2005; Senkowski et al. 2006). Integration of this information not only unifies the perception of events, but the presence of redundant information also facilitates recognition, increases signal-to-noise ratios, and decreases reaction times to crossmodal events (Driver and Spence 1998; Hershenson 1962; Senkowski et al. 2006; Stein et al. 1989). Studies examining the simultaneous serial and parallel computations and physiological responses underlying the integration of information and the recognition of a unified percept have important implications for advancing the understanding of the binding of crossmodal information for ecologically valid behaviors such as motion perception and speech recognition and comprehension (Baumann and Greenlee 2007; Lakatos et al. 2008; Miller and D'Esposito 2005; Schroeder and Lakatos 2009; Schroeder et al. 2008).

While it has traditionally been thought that processing of crossmodal events occurs primarily in association cortices (Jones and Powell 1970; Mesulam 1998),

much recent evidence indicates that information from other sensory modalities can influence cortical areas conventionally assumed to be unimodal (Ghazanfar and Schroeder 2006). For example, electroencephalographic (EEG), functional magnetic resonance imaging (fMRI) and magnetoencephalographic (MEG) studies in humans have provided evidence that visual and somatosensory signals can influence neuronal activity in the auditory cortex (e.g., see Schroeder and Foxe 2005 for a review). Intracranial recordings and anatomical tracings in macaques have affirmed the existence of multisensory inputs to putatively unimodal cortical areas (Kayser et al. 2008). In humans, several functional imaging and intracranial studies have identified cortical networks involved in object recognition, auditory-somatosensory and visual-somatosensory processing and integration of audiovisual speech (Calvert et al. 1999, 2000, 2001; Molholm et al. 2004, 2006; Senkowski et al. 2008). Human imaging studies have identified the superior colliculus, superior temporal sulcus, intraparietal sulcus, insula and several frontal cortical areas as being involved in crossmodal computation (Calvert et al. 2001). With regard to speech, the traditional speech areas (perisylvian) have been implicated, as well as the superior parietal, inferior parietal, inferior frontal, superior temporal sulcus and left claustrum areas (Calvert et al. 2000; Campbell 2008; Fort et al. 2002; Olson et al. 2002). These findings emphasize the importance of rapid synchronization of crossmodal information in heteromodal cortical areas.

A number of event-related potential (ERP) studies have examined the temporal aspects of crossmodal interactions, motivated by the hypothesis that the

decrease in reaction time and facilitation of object recognition should be visible in electrophysiological recordings. These studies have found significant activity within several latency windows, with the most surprising results for audiovisual interactions coming at ~50–100 ms post-stimulus onset, suggesting early cortical processing of audiovisual interactions (Molholm et al. 2002). In addition, several ERP studies have also evaluated facilitation of bimodal interactions via an additive model (Besle et al. 2004). These studies typically have shown amplitude and latency facilitation due to bimodal interactions localized to multimodal cortical areas, as well as suppression of electrophysiological responses with cortical generators in (putatively) unimodal areas.

A slightly different electrophysiological paradigm for investigating the computational advantages of crossmodal interactions is provided by the steady-state response (SSR), which is the result of entrainment to the temporal properties of a modulated stimulus. This response has been documented for both visual and auditory signals and has been used extensively for clinical and diagnostic purposes (Sohmer et al. 1977). Auditory SSRs are generally elicited by amplitude or frequency modulated signals, or both (e.g., Luo et al. 2006), while visual SSRs are typically elicited by transient high-contrast stimuli such as checkerboard reversals or luminance flicker. Though commonly measured with EEG, the same principles of frequency entrainment to periodic stimuli have been evaluated in MEG (Muller et al. 1997; Ross et al. 2000). Ecological stimuli that are temporally extended and have a quasi-steady-state nature, such as speech, can also be modeled via stimuli that approximate the excitation produced by domain-specific information (Grant and Seitz 2000). SSRs

have a potential further advantage: they can be used to exploit endogenous cortical oscillations. These oscillations are amplified when preferential stimuli (i.e., stimuli that match the frequency and phase of the endogenous oscillations) constitute the sensory input (Schroeder and Lakatos 2009; Schroeder et al. 2008; Senkowski et al. 2008). Oscillatory activity of particular interest occurs in frequency ranges that are important for relevant behaviors such as speech comprehension, working memory function and selectional attention (Luo et al. 2010; Luo and Poeppel 2007; Senkowski et al. 2008; Talsma et al. 2006).

The motivation for the current study was to model an ecologically typical audiovisual interaction, multisensory speech, incorporating some of its critical temporal attributes. The auditory component of speech consists of relatively rapid frequency fluctuations in the spectral domain, along with slower amplitude modulation (i.e., the envelope)—reminiscent of an amplitude-modulated (AM) sinusoidal auditory signal. The speech signal itself shows significant AM activity in the 2–16 Hz range (Steeneken and Houtgast 1980), and it has been shown that cortical decomposition of the speech envelope is particularly sensitive to frequencies in the range of 4–16 Hz. For example, recent MEG evidence supports this generalization: Luo and Poeppel (2007) and Howard and Poeppel (2010) observed that fluctuations in the speech envelope are associated with intrinsic oscillations in the theta frequency band (~4–8 Hz). Luo et al. (2010) extended that to the delta band (1–3 Hz) when AV speech was used. Paired with the auditory signal is a visual component in which facial features—and especially mouth movements—aid

comprehension, especially in noisy environments (Sumbly and Pollack 1954). We thus crafted stimuli consisting of modulated auditory and visual components within the frequency range of the envelope of speech. By building on results investigating SSRs to auditory and visual stimuli presented alone, we assess the SSR to bimodal audiovisual signals.

For the experiment reported, the visual signal consists of an ellipse to approximate a mouth opening and closing, and the auditory signal consists of amplitude-modulated three-octave pink noise to approximate the envelope and the wideband carrier features of the speech signal. We hypothesize that the SSRs elicited by concurrently modulated (comodal) audiovisual signals should be greater than the responses elicited by unimodally modulated auditory or visual stimuli, as reflected by the amplitude spectrum at the modulation frequency and its harmonics. The increased signal power of the comodal conditions relative to unimodal conditions might reflect increased activity due to synchrony of different neural populations involved in evaluating the multimodal signal. By manipulating the phase congruence of one modality relative to the other, we additionally aimed to elucidate the online cross-talk between modalities. In the experiment presented here, we demonstrate the feasibility of bimodal SSR as an experimental paradigm, the cortical/sensor areas that are predictors for evaluating signal component synchronicity, and the particular SSR component that indexes signal envelope component synchronicity.

Materials and Methods

The experimental design and data presented result from an earlier experimental design where we determined the feasibility of eliciting a bimodal SSR (at two distinct modulation frequencies) as well as the most appropriate methods to analyze the data. The major methodological points from that pilot experiment are as follows: (i) pretest stimuli to determine sensors that responded preferentially to a given modality (see pretests below); (ii) for the SSR analysis, we first averaged trial presentations, multiplied the data within the signal evaluation window in the temporal domain by a Kaiser window ($\beta = 13$) to remove spurious frequency contributions and minimize onset and offset responses and then evaluated the Fourier transform; (iii) evaluated the significance of SSR activity using a combination of F tests and Rayleigh's phase test; (iv) evaluated onset responses using principal component analysis (PCA); (v) performed a crossmodal control analysis to verify the responses recorded were not sensor-dependent and (vi) evaluated the significance of the results using a combination of parametric and non-parametric statistical tests.

The results of the prior experiment revealed whether or not there was any asymmetry in the response power to each modality, and also allowed us to fine-tune the signal evaluation methods and statistics used to evaluate the responses, which are applied to the data presented below.

Participants

Fourteen participants (thirteen right-handed; one ambidextrous, as tested by the Edinburgh Handedness Inventory (Oldfield 1971); six female) with normal hearing and normal or corrected-to-normal vision underwent MEG scanning. Data from two participants were excluded due to an insufficient signal-to-noise ratio for all conditions. Age range was 18–27 (mean 20.1 years). Participants were compensated for their participation (\$10/hr). Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant.

Stimuli

In the strictest sense, all signals presented were bimodal; we use the terms “unimodal” and “comodal” in a specific manner here to distinguish signal types. Unimodal refers to conditions where only one modality undergoes envelope modulation, while the other modality is presented with a static signal to control for basic sensory excitation. Comodal conditions refer to simultaneous envelope modulation of both modalities.

As mentioned in the *Introduction*, the modulation frequency employed was intended to model the naturalistic modulation rates found in speech. Though the majority of SSR studies in both the visual and auditory domains use high modulation rates (e.g., 30–80 Hz), we observed unique challenges most likely due to the shape

and contrast of the experimental signals and the extended duration of the experiment (approximately 60 min). During a piloting phase, we tested modulation frequencies of 6–8 Hz. Although we observed SSR activity at these frequencies, modulation at these rates caused participant eye fatigue and discomfort, leading us to lower the modulation frequency to 3.125 Hz. This modulation frequency satisfies several criteria: (i) it is in the range of speech modulation, (ii) it falls within a specific frequency bin to eliminate the need for filtering and windowing the SSR response to make signal extraction easier and reduce information loss and (iii) it was more comfortable for participants to attend to.

The SSR-inducing stimuli consisted of five types of audiovisual signals presented at one modulation frequency, along with three target signals, for a total of eight signals. The five types of SSR-inducing signals were: (i) amplitude-modulated three-octave pink noise presented concurrently with a static white rectangle on black background; (ii) a radius-modulated white ellipse on black background concurrently presented with approximately Gaussian white acoustic noise; (iii–v) a radius-modulated ellipse paired with amplitude modulated three-octave pink noise at one of three envelope phase relationships (in phase, $\pi/2$ radians out of phase, π radians out of phase). The amplitude-modulated three-octave pink noise and radius-modulated ellipses were modulated at 3.125 Hz with a modulation depth of 25% of peak amplitude and radius for audio and visual signals, respectively (Fig. 1). The SSR-inducing signals were 3.520 s in duration. For the comodulated conditions, the auditory and visual signal components had the same onset and offset, with the

auditory component reaching the maximum value of the modulation envelope first for out-of-phase conditions.

Auditory signal components were generated with MATLAB (R2009a, The Mathworks, Natick, MA) and consisted of a cosine wave envelope (3.125 Hz modulation frequency) applied to a three-octave pink noise carrier signal with 6 ms \cos^2 onset and offset ramps presented at approximately 65 dB SPL. Using the cosine function for the envelopes was done to enhance the onset responses; the cosine function starts at the maximum value, which will generate a more robust onset response. The pink noise was generated using the NSL Toolbox (Chi and Shamma, <http://www.isr.umd.edu/Labs/NSL/Software.htm>) for MATLAB. The three-octave pink noise contained a lowest frequency of 125 Hz; these parameters cover the fundamental frequency range of the human voice as well as the frequency region where most of the energy arising from the first formant tends to be concentrated. The signals were sampled at 44.1 kHz with 16-bit resolution. Visual signal components were generated using Gnu Image Manipulation Program (www.gimp.org). The radius-modulated white ellipses were centered on a 640 x 480 pixel black background, and ranged from 0.84° to 1.68° visual angle for the minor radius and 3.71° visual angle for the major radius. The minor radius was modulated to simulate mouth movements. The individual frames were compiled into Audio–Video Interleave (AVI) format using Virtual Dub (www.virtualdub.org) for presentation. Stimulus timing/frequency was verified with an oscilloscope. The visual components

were projected on a screen approximately 30 cm from the participant's nasion. Participants were supine in the MEG scanner for the duration of the experiment.

To maintain participant vigilance to both modalities, brief targets were pseudorandomly interleaved throughout the experimental trials. Targets were of three types: (i) an auditory only target consisting of approximately Gaussian white noise; (ii) a visual only target consisting of a white crosshair on a black background; (iii) an audiovisual target consisting of a white crosshair on black background paired with approximately Gaussian white noise. Target duration was 500 ms.

Experimental stimuli were presented in six blocks, with 15 repetitions per signal per block, for a total of ninety trials per condition. Presentation of conditions was randomized within blocks. The SSR-inducing materials were passively attended to; no response to those signals was required. For the target signals (38% of trials), participants were required to press a button indicating their detection of the target.

Delivery

All experimental stimuli were presented using a Dell Optiplex computer with a M-Audio Audiophile 2496 soundcard (Avid Technology, Inc., Irwindale, CA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Stimuli were delivered to the participants binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). The inter-stimulus interval varied pseudo-randomly between 980 and 2000 ms.

Recording and Filtering

Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was DC-200 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. The data were noise reduced using time-shifted PCA (de Cheveigné and Simon 2007) and trials were averaged offline (artifact rejection ± 2.5 pT) and baseline corrected. Data for the SSR analysis were not filtered; however, data for examining the onset responses were filtered. The filter employed was a 4th order lowpass elliptical filter with a 40 Hz cutoff frequency, 0.5 dB peak-to-peak ripple and at least 60 dB stopband attenuation.

Sensor Selection from Pre-Test

Determination of maximally responsive auditory and visual channels was performed in separate pre-tests. The auditory pre-test consisted of amplitude-modulated sinusoidal signals with an 800 Hz sinusoidal carrier signal, modulation frequency (Fm) 7 Hz, modulation depth 100% and 11.3 s duration. The visual pre-test consisted of a checkerboard flicker pattern (Fm = 4 Hz), of 240 s duration. The sensor space was divided into quadrants to characterize the auditory response and sextants to characterize the visual response based on the peak and trough field topography expected for each modality as recorded from axial gradiometers (see Fig. 2). Sensor channel designations were anterior temporal (front of head), posterior

temporal (rear quadrants/middle of head) and occipital (back of head overlying occipital lobe). Five channels from source and sink from each sensor division (i.e., ten channels for auditory response and five channels for visual response per hemisphere; 15 channels per hemisphere total) with the maximum measured magnetic field deflection were used for subsequent analyses. The analysis window for the PSD analysis of the visual pretest was 10 s and for the auditory pretest 11 s.

Onset Response Evaluation

The signal evaluation window (averaged and filtered sensor data) ranged from 500 ms pre-trigger to 3519 ms post-trigger. For several participants with exceptionally clean and robust onset responses, examination of the data revealed three distinct evoked peaks: (i) in the range of ~70–85 ms post-stimulus onset, with an auditory magnetic field topography; (ii) in the range of ~120–150 ms post-stimulus onset, with a visual field topography and (iii) in the range of ~180–240 ms post-stimulus onset, which was a combination of auditory and visual topographies. For the majority of participants however, such clear response patterns were not observed. Due to the univariate design, we calculated the RMS of the individual participant RMS vectors, and grand averages of the magnetic field deflections. Permutation tests were performed on time ranges obtained from visible peaks in the grand averaged waveform. Latency values used in permutation tests were taken from individual RMS vectors in the time ranges of the peaks observed. These data were taken from the filtered and baseline corrected individual participant data; baseline correction

within participants served as a normalization of the data used in the grand averages. The number of trials averaged was, at a minimum, 80 out of 90 presentations.

SSR Analysis

The magnitude and phase spectra of the SSR were determined using the Fast Fourier Transform (FFT) of the baseline corrected channel data. The FFT was calculated from 320 ms post-stimulus onset to the end of the signal evaluation window (3519 ms) for a total of 3200 samples analyzed; this yielded frequency bins commensurate with the modulation frequency and its harmonics. The magnitude of the response was calculated using the RMS of the FFT across channels. The phase response was determined by calculating the mean direction as described by Fisher (1996) based on the phase angle of the Fourier transformed data. The across participant response power was determined by calculating the mean of the individual participant power vectors. To determine the across participant phase response, the mean direction of the individual mean directions was calculated.

Across-Participant Response Averaging

Onset responses were collected and evaluated as described above. A similar procedure was used for the Fourier transformed data (collection of FFT vectors and grand averages computed). Individual participant vectors for response power (squared magnitude) and phase were collected and the relevant statistics calculated as described below.

Statistical Analyses

The significance of the SSR amplitude at a specific frequency was analyzed by performing an F test on the squared RMS (power) of the Fourier transformed data using the MATLAB Statistics Toolbox. The F test takes into account both amplitude and phase (Valdes et al. 1997; Picton et al. 2003). For the across-participant data, F tests were performed on the power of the SSR at the modulation frequency and the second harmonic. The response power in linear values and decibels (dB) was assessed using ANOVAs as well as General Linear Models (GLMs) using the “languageR” statistical package (R Foundation for Statistical Computing, v. 2.10.1; Baayen 2008). Factors for both sets of statistical tests were Hemisphere, Harmonic, Condition, and Sensor Area, with Participant as a random effect. To determine the separation of densities, distributions of the responses for each hemisphere, harmonic, condition and area were compared using Kolmogorov–Smirnov tests. Use of ANOVAs is standard when comparing responses across participants in electrophysiological experiments (Jenkins et al. 2010); we used the power afforded by GLMs to determine more robustly the predictors of the response. K–S tests were used to see if the response distributions in each of the sensor areas were statistically different.

Additionally, we compared response additivity using the AV versus (A + V) model, but not via RMS of the recorded responses (if the additivity were assessed using RMS, this would assume a single source is generating the response; because the

responses examined involve two sensory domains, this would not be parsimonious). As such, additivity evaluations were made using the complex representation from the Fourier transform of the data on the frequency bins containing the frequencies of interest, specifically the modulation frequency and the second harmonic. Responses at the third harmonic were not statistically different from background noise. Statistical differences were assessed using Wilcoxon signed-rank tests in order to decrease the assumptions concerning the distribution of the data recorded between pairs of conditions.

Participant Head Location

Though we did not perform dipole localization, we did take marker measurements and recorded digitized headshapes for the participants. The marker coils were placed by common anatomical markers: by each preauricular point and three frontal coils based on spacing from the nasion. Head position measurements were taken prior to and after experimental completion to determine proper head placement within the dewar and that the sensors were recording from the entire head (occipital, posterior temporal/parietal, anterior temporal/frontal areas). This also aided in ensuring that sensor selection from the pretests was correct.

Results

Figure 3 illustrates the nature of the data recorded using a single participant; magnetic flux is in black, RMS in red. The panels, from top to bottom, (i) data from pre-stimulus onset to the end of the analysis frame, (ii) a zoomed-in view of the onset

response and (iii) the steady-state portion of the evoked response. For the onset response, there are two distinct peaks in the RMS, one occurring at ~140 ms and the other at ~210 ms post-stimulus onset. The oscillatory activity seen in the evoked response indicates entrainment to the periodic properties of the experimental stimuli (verified via the Fourier transform of the magnetic signals and statistical assessment).

SSR responses were reliably generated. The response pattern observed indicated (as measured using ANOVAs and GLMs as well as data visualization) that there was no difference between hemispheres in the power of the response, and that the posterior temporal and occipital channels captured the response best (Fig. 4).

Examination and analysis of the SSR power indicated that it would be more advantageous to analyze the responses in terms of decibel (dB) power, rather than linear power values, due to the effectively normally distributed nature of dB power measurements (Dobie and Wilson 1996; see Fig. 5). Data visualization of power densities was performed using the “ggplot2” package for R (Wickham 2009). The dB values readily yield to a more robust and easily comprehensible statistical analysis.

Across-Participant Power Analysis

Most of the response power was generated in the sensors overlying the posterior temporal and occipital areas. Response power was concentrated at the modulation frequency and the second harmonic, and the power values at those

frequencies were used for the subsequent statistical analyses. Statistical significance was assessed using F tests with 2 and 12 degrees of freedom ($df = 2, 12, \alpha = 0.05$) and was confirmed by comparing the average power of the background noise (surrounding frequency bins) with the bin containing the modulation frequency. On average, the frequency bins containing the frequencies of interest were an order of magnitude (~ 10 dB) greater than the background, with exceptions for certain sensor areas and conditions (i.e., responses measured at anterior temporal sensors).

For the unimodal modulation conditions, statistically significant F ratios were found at the modulation frequency for the occipital sensors in both hemispheres (LH: $F = 37.441, p < 0.01$; RH: $F = 10.539, p < 0.01$), but not for the anterior and posterior temporal sensors; the second harmonic F ratio was significant only in the RH occipital sensors ($F = 7.853, p < 0.01$). For the $\Phi = 0$ comodula condition at the modulation frequency, significant F ratios were found for the posterior temporal and occipital sensors in the LH ($F = 7.822, p < 0.01$ and $F = 60.107, p < 0.01$, respectively); the RH occipital sensors F ratio was marginally significant ($F = 4.113, p < 0.05$); this same pattern held for the second harmonic ($F = 4.839, p < 0.05$; $F = 4.733, p < 0.05$; $F = 4.061, p < 0.05$, respectively). For the $\Phi = \pi/2$ condition, significant F ratios were found for the occipital sensors in both hemispheres at the modulation frequency (LH: $F = 74.436, p < 0.01$; RH: $F = 10.040, p < 0.01$) and the LH occipital sensors for the second harmonic ($F = 37.351, p < 0.01$). For the $\Phi = \pi$ condition, significant F ratios were found for the posterior temporal (LH: $F = 16.833, p < 0.01$; RH: $F = 7.358, p < 0.01$) and occipital sensors (LH: $F = 23.954, p < 0.01$;

RH: $F = 12.864$, $p < 0.01$) at the modulation frequency; at the second harmonic significant F ratios were found for the occipital sensors (LH: $F = 12.663$, $p < 0.01$; RH: $F = 8.127$, $p < 0.01$) and the RH posterior temporal sensors ($F = 3.901$, $p < 0.05$).

Statistical Summary

Separate ANOVAs were calculated with the following interactions: (i) Hemisphere (two levels) x Harmonic (two levels) x Condition (four levels) x Sensor Area (three levels), (ii) Harmonic x Condition x Sensor Area and (iii) Condition x Sensor Area. For the first ANOVA, significant interactions were found for Harmonic ($F(1,13) = 148.053$, $p < 0.001$), Sensor Area ($F(2,13) = 134.441$, $p < 0.001$), and Condition x Sensor Area ($F(6,13) = 4.208$, $p < 0.001$); the interaction Hemisphere x Sensor Area was marginally significant ($F(2,13) = 3.013$, $p = 0.049$). For the second ANOVA, significant interactions were found for Harmonic ($F(1,13) = 150.546$, $p < 0.001$), Sensor Area ($F(2,13) = 136.705$, $p < 0.001$) and Condition x Sensor Area ($F(6,13) = 4.279$, $p < 0.001$). For the third ANOVA, significant interactions were found for Sensor Area ($F(2,13) = 111.093$, $p < 0.001$) and Condition x Sensor Area ($F(6,13) = 3.477$, $p < 0.05$).

GLMs were then implemented to statistically determine the predictors of the responses (e.g., hemisphere, harmonic, condition, sensor area). GLMs used the same factors as the ANOVAs to evaluate the response power. For the first and second set of factors the second harmonic ($p < 0.05$), occipital sensors ($p < 0.01$), and the π

initial offset condition coupled with the posterior temporal sensors ($p < 0.05$) were predictors of the response power. For the third set of factors, the predictors were the posterior temporal sensors by themselves ($p < 0.05$), the occipital sensors ($p < 0.01$) and the posterior temporal sensors coupled with the three comodal conditions ($p < 0.05$).

Two-sample Kolmogorov–Smirnov tests indicated that the power distributions for the harmonics ($D = 0.324$, $p < 0.001$), anterior and posterior temporal sensors ($D = 0.455$, $p < 0.001$), anterior temporal and occipital sensors ($D = 0.4821$, $p < 0.001$) and posterior temporal and occipital sensors ($D = 0.134$, $p < 0.05$) differed significantly.

Post hoc analyses on the posterior temporal channels found significant interactions of Harmonic ($F(1,13) = 49.199$, $p < 0.001$; $F(1,13) = 50.157$, $p < 0.001$) and Condition ($F(3,13) = 10.103$, $p < 0.001$; $F(3,13) = 10.300$, $p < 0.001$) for the triple- and double-factor ANOVAs and Condition ($F(3,13) = 8.348$, $p < 0.001$) for the single-factor ANOVA. GLMs indicated statistically different response power predicted by the second harmonic (triple-factor: $p < 0.05$; double-factor: $p < 0.001$) and the comodal conditions (triple-factor: $p < 0.05$; double- and single-factor: $p < 0.001$).

SSR Power Comparisons

Figure 6 illustrates the differences in overall power between harmonics for each condition for the entire dataset for all sensor divisions (collapsed across hemispheres since there was no statistical difference in power between the hemispheres). Plots of the mean dB power show there is no statistical difference in power between the different conditions, but there is a difference in the power between harmonics, with the modulation frequency exhibiting greater power for each condition than the second harmonic. Additionally, though there is no statistical difference between conditions, the relational pattern of topographies observed seems commensurate with the hypotheses regarding representation of the comodal signal either as complete or separate entities (see *Discussion*).

Figure 7 illustrates the changes in response power for the posterior temporal (left panel) and occipital (right panel) sensors. Several trends can be observed. First, there is greater power at the modulation frequency than at the second harmonic. Second, the comodal conditions exhibit greater power than the unimodal conditions. Third, and most importantly, the difference in power between unimodal and comodal conditions seems to be directly attributable to the sensors overlying the posterior temporal areas (and possibly parietal lobes). No difference in power for either harmonic across conditions is observed in the occipital sensors.

Results of the Wilcoxon signed-rank tests for additivity indicated that the medians for the unimodal and comodal conditions did not differ except for $\Phi = \pi$ and

unimodal modulation pairwise comparison (LH: signed-rank = 3, $Z = -3.107$; RH: signed-rank = 4, $Z = -3.045$). This difference may be due to the nature of the representation used, as the Wilcoxon signed-rank tests for additivity used the complex numbers derived from the Fourier transform and not the power values as were used in the ANOVAs and GLMs.

Figures 8 and 9 illustrate the grand average topography at the modulation frequency and the second harmonic, respectively, in the form of phasor plots, which show the sink-source distribution and the phase of the response (Simon and Wang 2005). Two clear source-sink patterns can be observed for each frequency, while for each comodality condition more complex patterns are observed, especially for the second harmonic. The sink-source distribution (and phase distribution) at the modulation frequency (Fig. 8) for all conditions resembles that of a visual response recorded from axial gradiometer sensors; this is in line with the results from the power analyses, namely that the occipital sensors generated larger responses than the anterior and posterior temporal sensors.

For the response at the second harmonic (Fig. 9), the topographies seen are more complex, as they seem to reflect the degree of AV integration. For the unimodal auditory condition, the sink-source distribution reflects responses typically recorded from auditory cortex. For the unimodal visual condition, the sink-source distribution appears to be somewhat mixed. The sink-source distribution for the comodality conditions indicates (i) the degree of synchronicity and integration between

the signal components and (ii) the contribution of the posterior temporal sensors (and perhaps the auditory cortex and/or parietal lobes). For the $\Phi = 0$ condition, a clear auditory sink-source distribution is observed. For the $\Phi = \pi/2$ and $\Phi = \pi$ conditions, especially for the sensors overlying the posterior of the participants' heads, the sink-source distribution reflects the posterior auditory field topography, while for the remaining sensors the magnetic field distribution is not easily interpretable. Taken with the results of the statistical analyses, it is compelling that the changes in the response topographies and response power are due to the second harmonic and information from the posterior temporal lobes and/or auditory cortex, and possibly parietal lobes (Howard and Poeppel 2010).

Grand-averaged data yielded two peaks in the RMS for both unimodal modulation conditions and the comodal conditions. For the unimodal auditory condition these peaks occurred at ~ 140 and ~ 215 ms post-stimulus onset, for the unimodal visual condition the peaks occurred at ~ 109 and ~ 204 ms. For the comodal conditions, these peak latencies were ~ 140 ms for the first peak and ~ 210 – 217 ms for the second peak. These values suggested that synchronicity was also reflected in the temporal domain, because the peak latencies to comodal conditions were very close to those observed for unimodal auditory modulation and because the statistics on the SSR power indicated significant auditory contribution to the bimodal responses. However, the permutation tests did not show significant differences in peak latencies between conditions.

Discussion

The audiovisual MEG experiment presented has (i) extended a paradigm previously used to evaluate unimodal responses to investigate bimodal responses, (ii) elicited the bimodal SSR using novel stimulus types and (iii) elucidated some of the factors affecting the neural signal recorded. In the larger context of AV experiments, we have replicated several findings: (i) that visual contribution is greater than auditory (in the sense that the response power in visual areas is greater than in auditory areas) and (ii) when change is induced in bimodal signal components, the response in sensors overlying auditory areas changes the most, suggesting that auditory information contributes greatly when comodulated AV signals are presented, in particular when stimuli are temporally aligned across modalities.

Although this experiment contained a large number of trials and hence a high SNR, we did not find any differences between the three comododal conditions as we initially hypothesized; however, a potential pattern of integration is borne out in the phasor plots. The data we present show effects of condition reflected in the power of the second harmonic for particular sensor areas, suggesting long-term dynamics are reflected in the first two harmonics of the SSR. While we found no statistically significant increase in signal power overall (see Fig. 6), there was a significant increase in power at the second harmonic for the comododal signals relative to the unimodal signals. As illustrated in Figs. 8 and 9, which show the magnetic field topographies for each harmonic in phasor plots, there is a clear difference in source-sink distribution for each harmonic. At the modulation frequency, the source-sink

distribution mirrors that of a visual response, while at the second harmonic, the distribution observed mirrors that of an auditory response, depending on condition.

These topographic phasor plots (and the statistical results) suggest that the harmonics may be representing differential processing within and across modalities. The activity at the modulation frequency may reflect the modality where attention is directed or which is more salient to the observer (Talsma et al. 2006; Saupe et al. 2009; Gander et al. 2010). Second harmonic activity may reflect envelope congruency changes between modalities, which, based on the observed field patterns, may be related to the degree of statistical regularity and synchronicity in the overall signal. This response, most likely originating from auditory or parietal areas, contributes most to the neurocomputational analysis of comodular AV signals.

As mentioned previously, we did not have access to structural MRs for our participants. However, we are fairly certain of the cortical areas that are most likely generating these signals. First, the lack of MRs did not prevent getting an estimate of participant headshape (see *Materials and Methods*). As such, we had a reliable estimate of the shape and outline of each participant's head. Second, we were able to place the location of each participant's head within the scanner by using head marker coils; this allowed us to make sure all cortical areas of interest (based on headshape measurements) were being recorded. Lastly, special care was taken to select the sensors from domain-specific pretests. Combined, these procedures, though without

anatomical constraints, assisted us in narrowing down the most likely generators of the responses observed.

Prior to executing the experiment, we had hypothesized that the signals would be represented cortically in three ways. When the signal envelopes are completely congruent, the signals may be observed and ‘computed’ as a single object. When the initial envelope phase offset is $\pi/2$ radians, then over the time course of the comodular signal, the signal components would be alternately perceived as one or two objects, as the synchronicity changes between being out-of-phase and in-phase. Lastly, when the offset is π radians between component envelopes, then each component would be perceived as a single object. As the signal component envelopes are desynchronized, the correlations and redundancies in the bimodal signal decrease, modifying the processing and representation of the percept. To verify these hypotheses, a psychophysical task would have to be incorporated along with characterization of the electrophysiological responses. Although we have anecdotal data from experimental piloting and participant debriefing, the current data do not support these hypotheses.

The congruency potentially indexed by the phase separation in this paradigm may have practical limits. There is evidence that integration of bimodal signals, with the auditory signal leading, takes place within a 40–60 ms duration window (van Wassenhove et al. 2007). For the modulation frequencies employed here, the incongruity between signal components did not fall within this integration window. It is entirely possible that the response patterns we observe are dependent on the

modulation frequency. Higher envelope modulation rates (e.g., 7–11 Hz) with phase separations falling within the temporal window of AV integration could test the SSR response to perceptually simultaneous but physically asynchronous signals.

A related issue is to sample more phase separation values around the entire unit circle. One possible hypothesis is that the representation of the phase separation will be symmetric (except when both signal envelopes are completely synchronized), i.e. the response power for a phase separation of $\pi/2$ radians and $3\pi/2$ radians will be represented equally. The indexing of signal component congruity might also be dependent on which component reaches the maximum of the envelope first. It has been shown that when visual information precedes auditory information, signal detection and comprehension increases (Senkowski et al. 2007; van Wassenhove et al. 2007). In the current study, for the asynchronous bimodal conditions, the auditory component of the signal reached the maximum of the modulation envelope first. It would be useful to examine the interactions that occur when the visual component modulation envelope reaches the maximum value before the auditory envelope.

Adding ‘jitter’ or noise to the signal component envelopes may also yield a more ecologically valid set of stimuli for further experimentation. This would add the variability inherent in speech, while retaining the modulation information of the signal component envelopes. Finally, the modulation depth of the auditory signal component might be made more variable to correspond with the conditions occurring

in natural human speech, where the mouth opens and closes fully (modulation depth ranging from 0 to 100%).

Much in the same way as traditional unimodal steady state responses are used to probe auditory and visual function, it may be possible to use the paradigm we introduce to assess audiovisual integration in humans. Deviations from the 40 or 80 Hz aSSR response have been suggested to correlate with impairments in CN VIII, the brainstem, or possibly cortical processing (Valdes et al.1997). Application of this paradigm could be used as a clinical assessment of audiovisual integration.

In summary, we demonstrate that an experimental technique commonly applied to unimodal signals, the SSR, can be applied to signals of a bimodal nature approximating the spectro-temporal properties of speech. We observed that the presence of bimodal information increased response strength in auditory areas.

Our findings are in line with several studies regarding AV integration, especially with regard to the specific contributions of auditory information (Cappe et al. 2010). In a real-world stimulus analogous to our ‘noislipses’, Chandrasekaran et al. (2009) characterized bimodal speech stimuli (with no phase incongruities) and observed (i) a temporal correspondence between mouth opening and the auditory signal component envelope and (ii) mouth openings and vocal envelopes are modulated in the 2–7 Hz frequency range. That modulation frequencies in this range play a key neurophysiological role in the parsing of neurophysiological signals has

now been amply demonstrated. For example, Luo et al. (2010) show that audiovisual movies incorporating conversational speech bear a unique signature in the delta and theta neural response bands, values congruent with the Chandrasekaran behavioral data. Cumulatively, it is now rather clear that low-modulation frequency information lies at the basis of analyzing uni- and multimodal signals that have extended temporal structure. The results of this MEG study offer additional support for this claim, and future iterations of this paradigm could further elucidate the neural computations underlying multisensory perception of ecologically relevant stimuli.

Figures

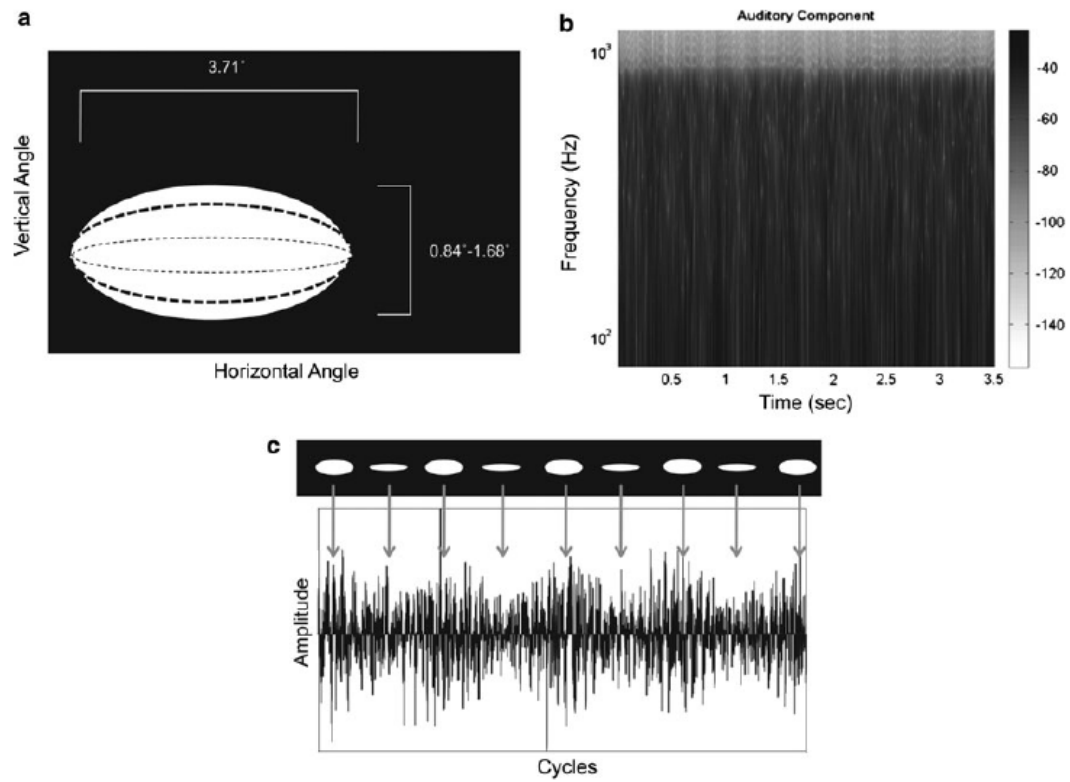


Figure 1a. Schematic of stimuli employed. This panel illustrates the movement of the visual signal component throughout the duration of stimulus (see *Materials and Methods* for details). The stimuli were presented at $F_m = 3.125$ Hz, modulation depth was 25%. The x-axis shows the horizontal visual angle of the signal component and the y-axis the ranges of vertical visual angles. Figure 1b. This panel illustrates the auditory signal component spectral structure. This component consisted of three-octave pink noise (lowest frequency: 125 Hz), amplitude modulated at 3.125 Hz, 25% modulation depth. Intensity values are plotted using a grayscale intensity axis. The x-axis is time in seconds; the y-axis is Frequency (Hz) on a logarithmic scale. Figure 1c. Schematized temporal evolution of comodul signal, signal component envelopes completely in phase, over several cycles. Top portion illustrates the visual

component; bottom portion the auditory component. The visual component is seen to ‘open’ and ‘close’ over the duration of the signal; it is easily seen that in this experimental condition, the ellipse modulation is synchronized with the amplitude modulations in the auditory component.

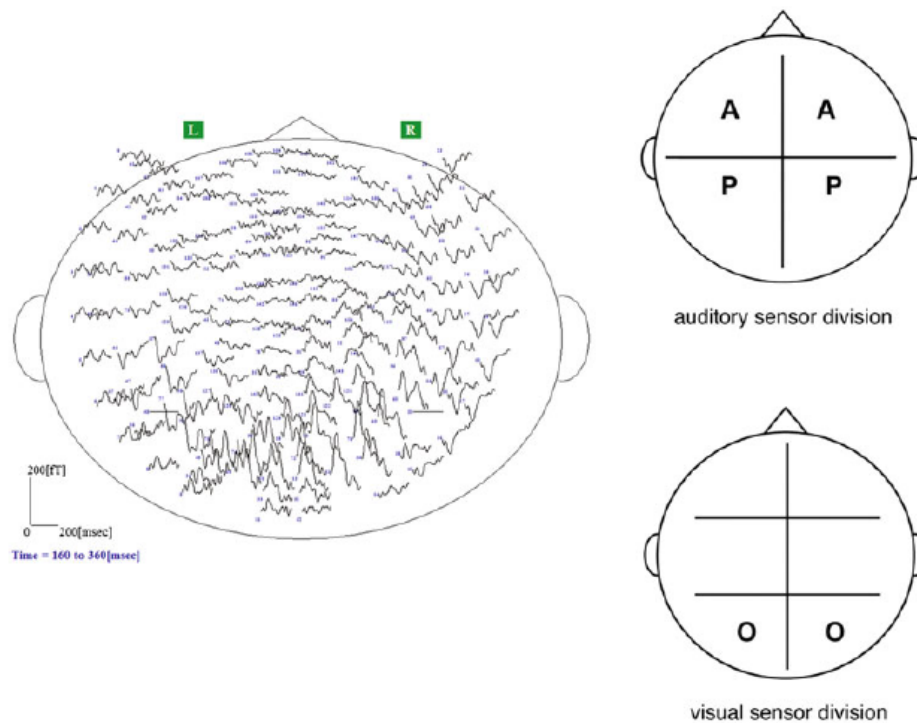


Figure 2. Left panel: Sensor layout of whole-head biomagnetometer with field deflections overlaid. Anterior portion of the head is at the top. This figure gives a sense of the positioning of the sensors in the dewar; robust evoked activity is seen at the channels overlying the occipital and posterior temporal lobes. Data are from a single participant, $\Phi = 0$ comodula condition. Right panel: Division of magnetoencephalographic sensors. Top panel shows division of auditory sensors for experimental pre-test; bottom panel shows sensor division for visual pre-test. Sensor

division was based on expected field topography for auditory and visual cortical responses recorded from axial gradiometer sensors (see *Materials and Methods* for details). Sensor designation is as follows: A = anterior temporal sensors, P = posterior temporal sensors, O = occipital sensors. Placement of letters roughly corresponds to the locations of the sensors selected for the analysis of the experimental data.

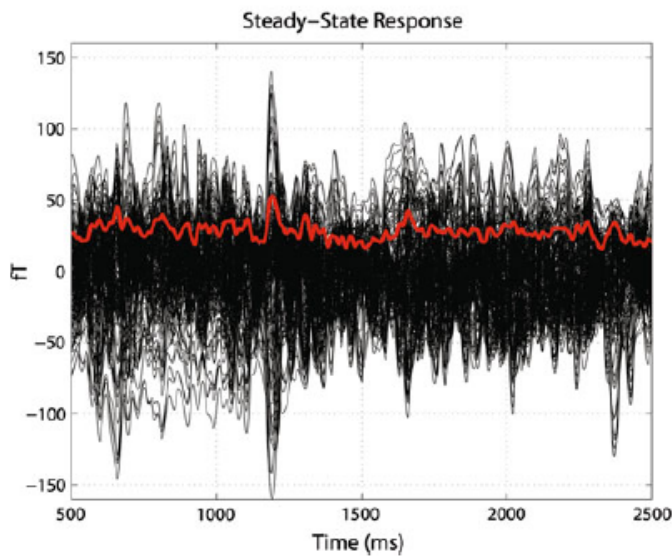
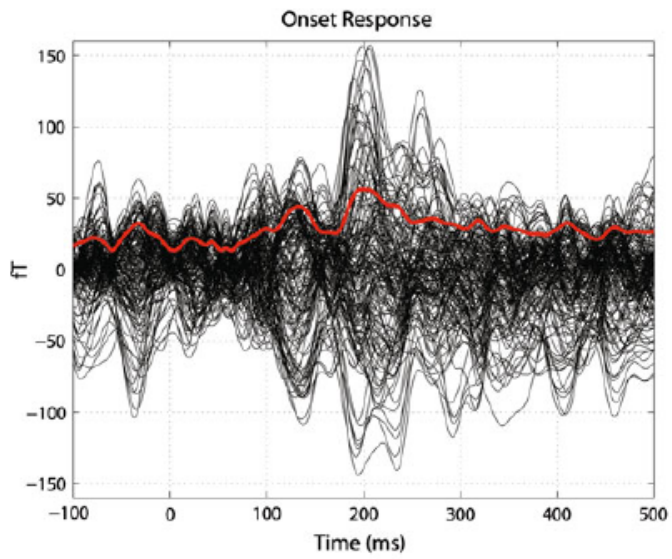
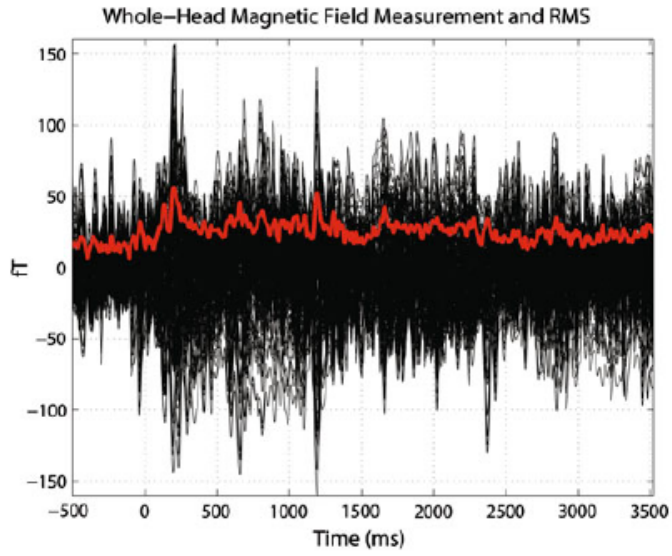


Figure 3. Top panel: MEG waveforms and RMS in the temporal domain from a single participant, $\Phi = 0$ comodal condition. Time is plotted on the x-axis; field deflection in fT on the y-axis. The top panel illustrates the magnetic fields recorded from all 157 data channels. Magnetic fields are in black, RMS is in red. The time duration shown is from 500 ms pre-stimulus onset to the end of the analysis frame. This panel illustrates both the onset (~ 0 –500 ms) and the steady-state response (towards end of onset to end of analysis frame). Middle panel: The middle panel provides a more detailed view of the onset response. Conventions are identical to the previous panel. Two clear peaks can be observed in the onset response at ~ 140 ms and ~ 210 ms post-stimulus onset (see *Results* for details). Bottom panel: Bottom panel illustrates the steady-state portion of the magnetic fields recorded. Conventions are the same as in the previous two panels. Oscillatory activity is clearly observed in the RMS of the signal, indicating entrainment to the physical structure of the comodal signals.

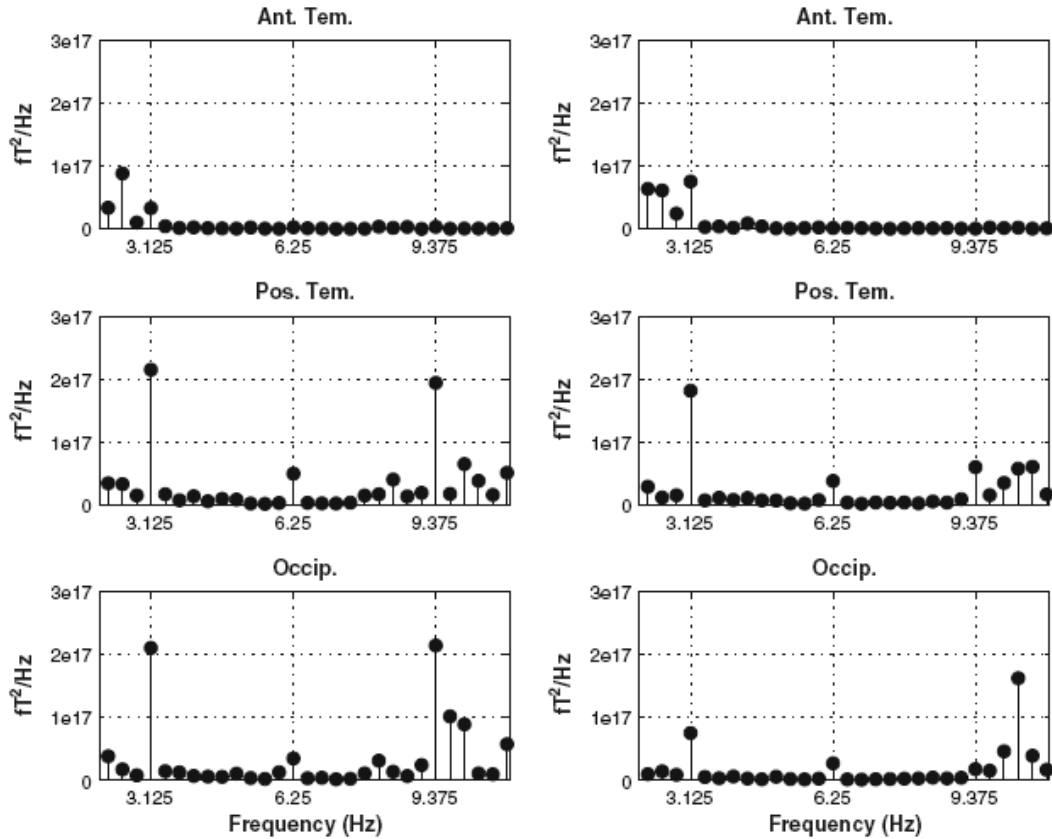


Figure 4. Grand averaged squared RMS power (linear) for all participants, completely synchronous comodal condition. Left column shows left hemisphere response, right column shows right hemisphere. Rows (top to bottom) show anterior temporal, posterior temporal, and occipital sensors. Hash marks on the x-axis indicate modulation frequency and second harmonic. It is clear that there is significant activity in the posterior temporal and occipital sensors at the frequencies of interest.

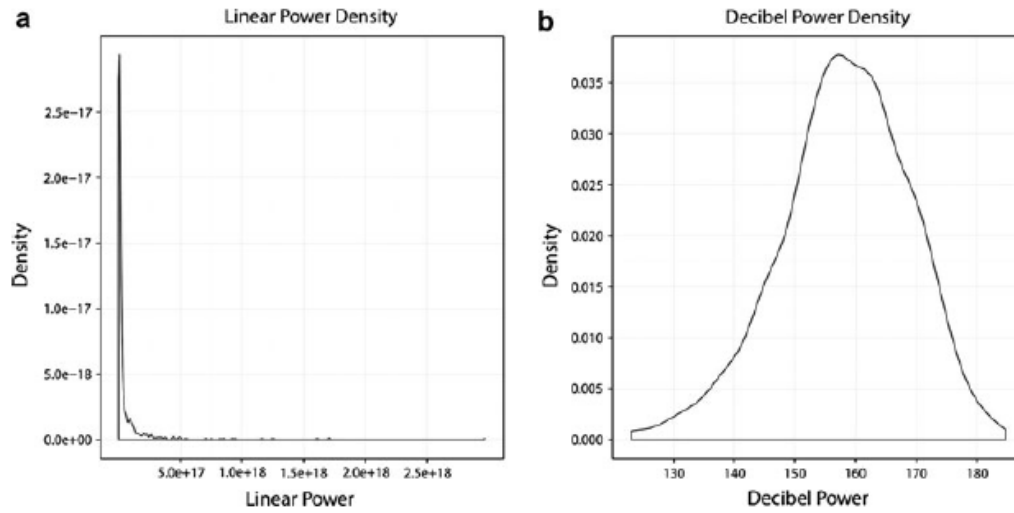


Figure 5. Density plots for linear (a) and decibel (b) power values. Linear power values are heavily skewed to the right and the combination of large numeric values and the skewedness of the distribution make these data somewhat hard to interpret visually and statistically. Subsequent analyses focus on decibel power values. Though still somewhat skewed, the decibel power values are more normally distributed than the linear power values, which yields to more easily interpretable visualization and statistical analysis. Additionally, representation and analysis of the data in this manner has been previously performed in the literature (Dobie and Wilson 1996) and may be more biologically plausible.

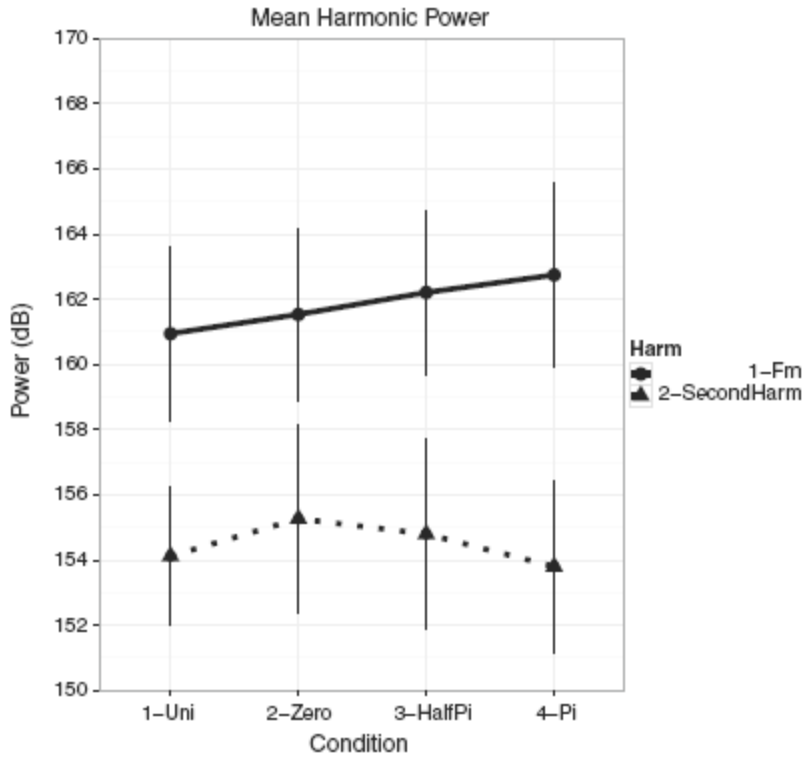


Figure 6. Mean harmonic power at the modulation frequency (3.125 Hz) and second harmonic (6.250 Hz); experimental condition is on the abscissa and power (dB) is on the ordinate. Power is collapsed across hemispheres and sensor areas. Solid line with circles denotes the modulation frequency and dotted line with triangles denotes the second harmonic. While there is no statistical difference in power between conditions, there is a clear separation in power between harmonics, with the second harmonic exhibiting lower power values than the modulation frequency, a result typical of SSRs.

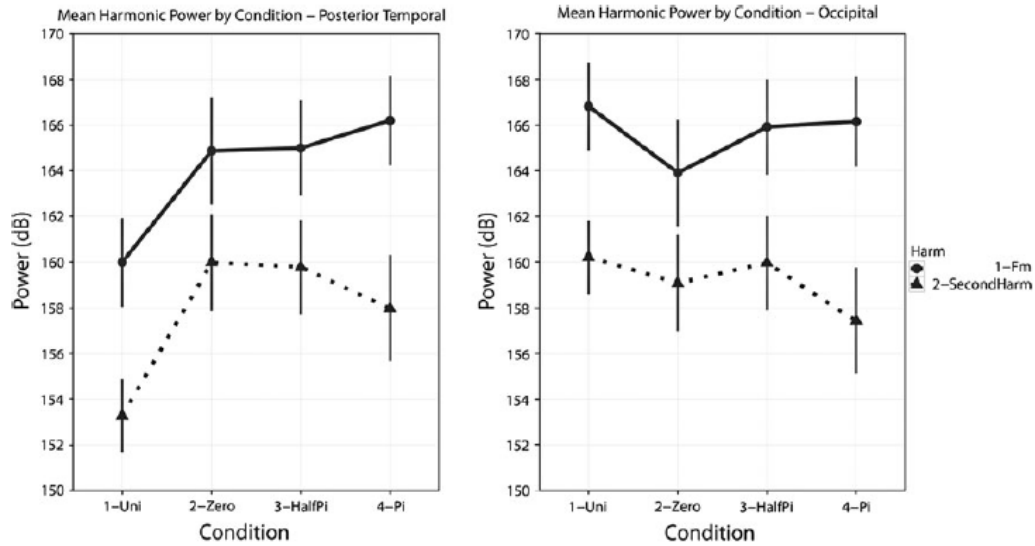


Figure 7. Left panel: Mean harmonic power for the modulation frequency and second harmonic by experimental condition for the posterior temporal sensors. Conventions used are the same as in Figure 6. Several trends can be observed: (i) response power at the second harmonic is lower than at the modulation frequency; (ii) the power for all three comodal conditions is greater than the unimodal conditions and (iii) there is no overall power difference between the comodal conditions. Right panel: Mean harmonic power for the modulation frequency and second harmonic by experimental condition for the occipital sensors. For the occipital sensors, the power at the modulation frequency is greater than that at the second harmonic and the power for all conditions in the occipital sensors is greater than that of the posterior temporal sensors. For the occipital sensors, there is no statistical difference between either the unimodal or comodal conditions or the comodal conditions themselves.

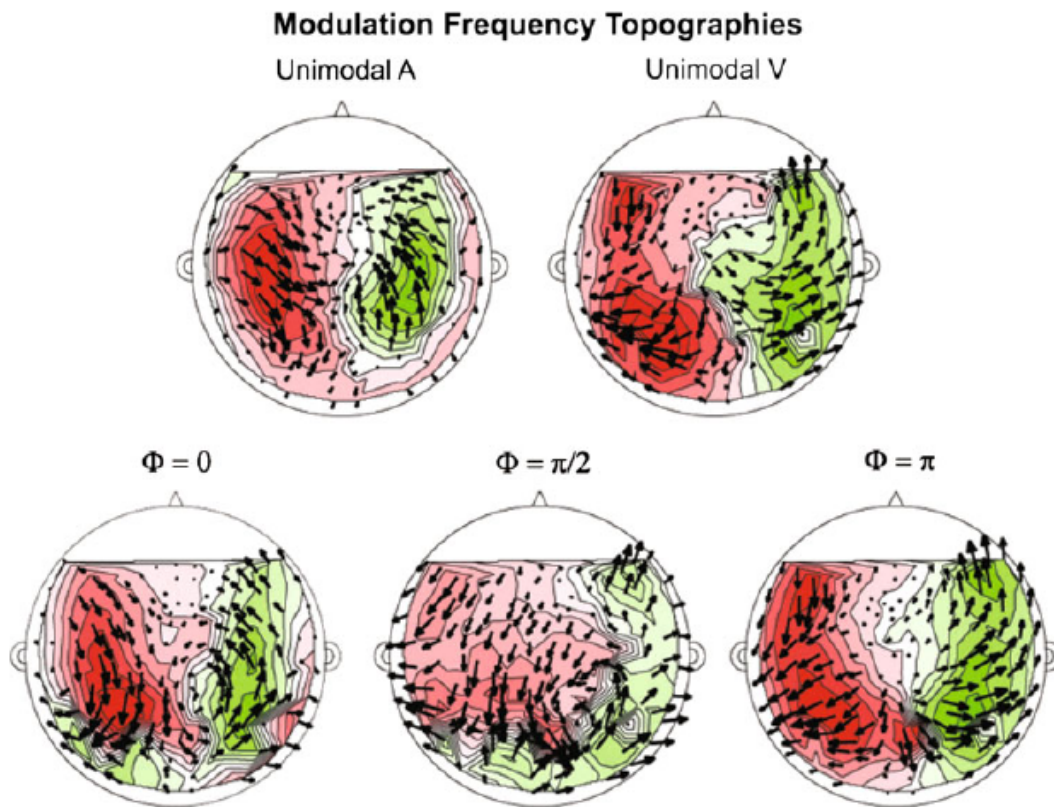


Figure 8. Phasor plot of grand averaged complex-valued topography at the modulation frequency (3.125 Hz) for each experimental condition. Top row shows unimodal conditions, bottom row comodal conditions. Magnetic source is indicated by green and magnetic sink by red. Phasors (arrows) indicate overall phase coherence and direction. For all experimental conditions, the source-sink distribution observed resembles that of a visual response as recorded by axial gradiometers. The topographies observed are in accordance with the finding that the overall visual response is greater than the auditory response, even for unimodal auditory modulation.

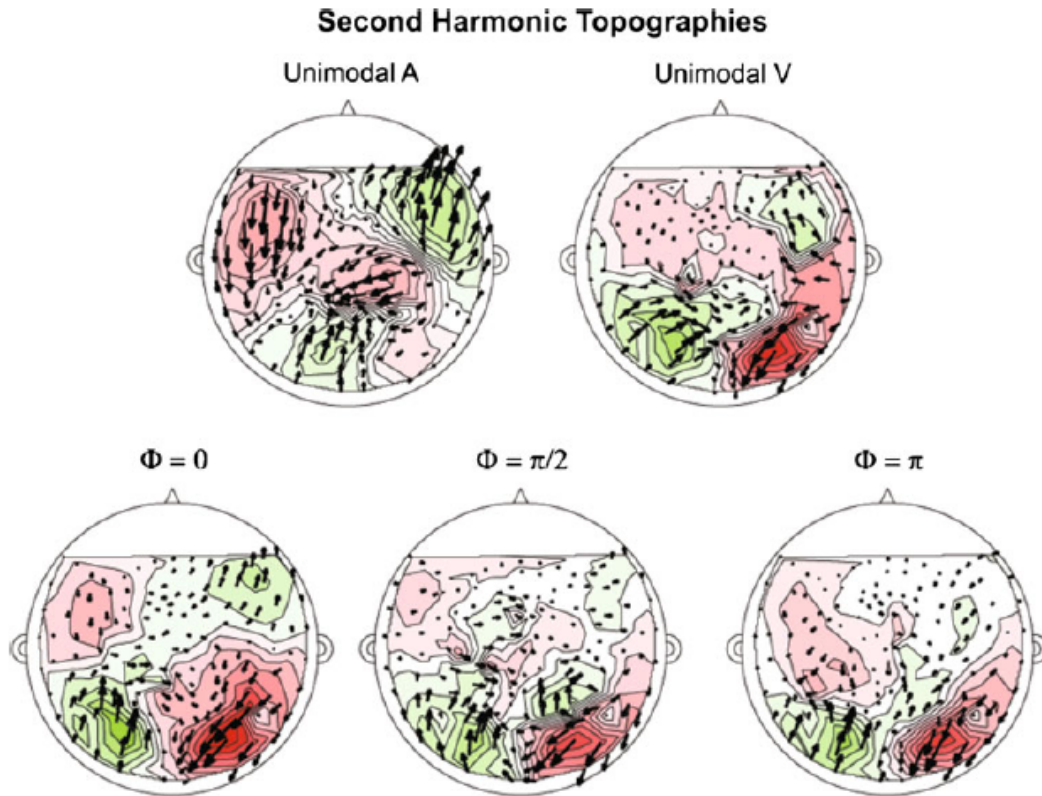


Figure 9. Phasor plot of grand-averaged complex-valued topography at the second harmonic (6.250 Hz) for each experimental condition. Conventions used are the same as in Figure 8. The source-sink distributions observed at the second harmonic more closely resemble that of an auditory response as recorded by axial gradiometers. For unimodal auditory modulation, the pattern observed (sink-source distribution and phasor direction and distribution) are rather clear, while for unimodal visual modulation, the topography is muted somewhat, but still observable. The source-sink distribution changes most significantly for the comodal conditions. For the $\Phi = 0$ condition, a clear auditory pattern is observed, while for the $\Phi = \pi/2$ and $\Phi = \pi$ conditions the topography seems to be a mix of auditory and visual activation. The constant between the three comodal conditions is that the sink-source distributions

towards the posterior end of the sensor distribution (posterior temporal and occipital sensors) resembles an auditory response topography. These plots agree well with the statistical results that the changes between conditions are predicted by the second harmonic, specifically in the posterior temporal sensors.

Psychophysical Discrimination and Clustering of Ecologically Approximate Synthetic Signals

Introduction

Our current understanding of how the auditory system creates functional groupings of auditory signals is underspecified. The prevailing and most intuitive viewpoint focuses on the physical characteristics of the signal (e.g., temporal and spectral features) being primary for the classification of the signals encountered in the environment. Neuroimaging methodologies, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) have identified cortical areas that specialize in the recognition of auditory objects (Binder et al., 2004) as well as selectively discriminating speech and vocal information (Belin et al., 2000; Belin et al., 2004; Hickok and Poeppel, 2007). The auditory percept most responsible for these processes is timbre. Timbre (or tone color) is an extremely salient multidimensional percept associated with the spectro-temporal profile of a signal. Timbre is “considered to be the acoustic quality of auditory objects [signals] and is integral to the identity of the source” (McAdams et al., 1995; Moore, 1995; Lakatos et al., 1997; Lakatos, 2000; Meyer et al., 2006). Timbre then, is a “psychologically relevant” auditory cue whose physiological and perceptual bases are not currently well understood (Meyer et al., 2006).

Due to its multidimensional nature, timbre perception is somewhat difficult to study. The multidimensional nature of timbre arises from spectral information, rise time between signal segments (frequency modulation -- FM) and envelope (amplitude modulation -- AM) evolution. Further complicating the investigation of timbre perception is the fact that auditory attributes, such as pitch, loudness and spectral content are not separable to some extent (Melara and Marks, 1990b, a) However, timbre studies have converged on a set of common (mostly orthogonal, i.e. independent) dimensions (McAdams et al., 1995). These dimensions commonly include attack time (rise time of envelope), spectral centroid (amplitude weighted frequency content), spectral flux (spectro-temporal evolution), spectral spread (bandwidth and spacing), spectral irregularity (harmonicity and inharmonicity), spectrum fine structure, and 'brightness' (odd and even harmonics contained in the spectrum) (Plomp et al., 1970; von Bismarck, 1974b, a; Grey, 1977; Hall, 1977; Kendall and Carterette, 1993; McAdams et al., 1995; Samson et al., 1997; Pressnitzer and McAdams, 1999; Pressnitzer et al., 2000; Marozeau et al., 2003; Caclin et al., 2005; Caclin et al., 2006; Caclin et al., 2007; Caclin et al., 2008). Based on observations of correlations between acoustic structure and behavior, it has been possible to design experiments that yield estimates of the perceptual space as well as test the interactions between the dimensions commonly reported (Samson et al., 1997; Lakatos, 2000). Signals of special interest for probing the structure of perceptual space and the interactions between acoustic signal structure and cognitive perception and processing are metamers -- signals that have different physical structures, yet are perceived as being the same. Probing the physical bases of perceptual resolution

(e.g., frequency, temporal, spectro-temporal integration) will help advance understanding of how functional categories are created and accessed.

Timbre has been extensively studied by using such an approach by employing a variety of signals, such as parametrically controlled synthesized signals, music, vowel tokens and musical tones and sequences, with signal space estimations recovered from behavioral output. Paradigms that have been utilized include verbal ratings and semantic scales (von Bismarck, 1974b, a), numerical scales (Slawson, 1968), multidimensional scaling (MDS) (Pols et al., 1969; Caclin et al., 2005) and Garner interference (Caclin et al., 2007). MDS studies have been crucial in elucidating the dimensions that are most salient to timbre perception, while ratings and scales have been fundamental in providing cognitive descriptions of signal attributes. Typically, two to four dimensions have been consistently found in psychophysical studies examining timbre perception; these dimensions tend to be orthogonal (independent), though there is some evidence that nonorthogonal factors are used in judgments (Lakatos, 2000). Additionally, when the studies employ vocal or vocal-based experimental signals, human observers may or may not use linguistic information in distinguishing or categorizing signals (Sinnott et al., 1997).

With respect to speech and voice perception, timbre contributes to such diverse processes as speaker identification, gender and affective state (Fellowes et al., 1997; Belin et al., 2000; Binder et al., 2004). In regards to the ecological speech signal, the excitation of the vocal folds and speaker vocal tract shape determines the

timbre of segments produced (generally phonemes). However, while there is a robust literature on speech production and comprehension, the multidimensional nature of the speech signal itself confounds investigation to some degree and the entirety of the complexity inherent in the signal is beyond our current understanding (Cleveland, 1977; Chartrand and Belin, 2006). To that end, in the experiment presented in this paper, we use synthetic signals that share some aspects of the speech and vocal signals but also allow us to probe timbre perception in a more general sense.

To be more specific, we model the production of acoustic signals in general and speech signals specifically in accordance with source-filter theory (Fant, 1980; Lakatos et al., 1997; Fant et al., 2000; Obleser et al., 2003; Tiitinen et al., 2005; Diehl, 2008). Here, a source signal is produced by an object, and a filter shapes the source signal by providing frequency-specific attenuation and boosting, giving rise to its timbre. In fact, there is evidence that source information can be extracted and exploited in characterizing the perceptual attributes of signals encountered (Slawson, 1968; Lakatos et al., 1997). In the case of a speech signal, the source waveform is the excitation pattern produced by the vocal folds and the shape of the vocal tract acts as the filter. For example, vowels are all-pole filters (Stevens, 1998), and the relationship of the poles (formant values) gives each vowel their unique timbre (e.g., formant transitions -- spectro-temporal evolution, spectral peaks -- pitch, spectral center of gravity/centroid and fine structure information -- glottal excitation). Speech (or more generally, vocal) signals also have a typical structure: they tend to be harmonic (Lewicki, 2002; Tiitinen et al., 2004, 2005), and the auditory system

decomposes the signal in such a way that ecological signals are perhaps preferentially and efficiently encoded (Sinnott et al., 1997; Lewicki, 2002; Smith and Lewicki, 2006; Diehl, 2008). The ability to preferentially encode and learn specific signal structures is also present at very early ages and may be important for the learning of linguistic patterns (Diehl, 1981; Thorpe et al., 1988; Trehub et al., 1990; Clarkson and Clifton, 1995; Trainor et al., 2004; Werker and Yeung, 2005; Cristia and Seidl, 2008).

In regards to the perception and processing of vowels (natural or synthesized), neuroimaging and behavioral studies have shown that several features are important for the physiological and perceptual analysis of vowel signals. For example, the first two formant frequencies seem to be crucial to vowel categorization and appear to be related to the first two principal components extracted from behavioral information (Slawson, 1968; Pols et al., 1969; Vihla et al., 2000; Jacobsen et al., 2004b; Tiitinen et al., 2004, 2005). Other factors crucial to vowel perception include general spectral features (Eulitz et al., 1995; Ragot and Lepaul-Ercole, 1996; Diesch and Luce, 1997; Poeppel et al., 1997), speaker information (Ragot and Lepaul-Ercole, 1996; Poeppel et al., 1997), harmonic structure (Darwin and Gardner, 1986) and phonetic and articulatory information (Tiitinen et al., 2005). Results of neuroimaging experiments link physiological processing propagated from the periphery to cortical and behavioral measures of timbre perception.

The psychophysical experiment presented here uses four different source signals and three different vowel spectra as filters. The vowel spectra are derived

from American English male formant measurements /a/, /i/ and /u/ (Hillenbrand et al., 1995) and the source waveforms are a sawtooth wave comprised of twenty-three harmonics, a sawtooth wave with every fourth harmonic removed, a sawtooth wave with every second harmonic removed and a square wave where the maximum harmonic is the twenty-third. Even though the signals presented are entirely synthetic approximations to speech signals, the individual vowel transfer functions are intended to act as distinct categories, based on phonetic and spectral information (Pastore et al., 1990; Jacobsen et al., 2004a; Jacobsen et al., 2004b). Removal of harmonics creates a timbral continuum (Singh and Hirsh, 1992; Jones and Perez, 2001) and the sawtooth wave with every second harmonic removed and the square wave are metamers (there is a slight phase difference between the signals).

We employ a one-alternative (interval)-forced choice (1AFC/1IFC) same-different (or AX discrimination) experiment in order to determine how sounds of differing and identical timbres are perceived. We employed this paradigm due to its ease and efficiency and due to the fact that the participants did not need to know *a priori* on what basis the signals differed. In addition, we also conducted a simple modeling study using various representations of a broader timbre continuum: the physical structure of the waveform, the power spectral density (PSD) and a cochlear filterbank representation, clustered using k-means clustering, hierarchical agglomerative clustering and scaling based on dissimilarity measures (Fellows et al., 1997; Lakatos, 2000). The goal of the psychophysical experiment was to (i) probe timbre sensitivity, (ii) test the resolution of timbre perception via metamers and (iii)

see how acoustic structure and linguistic information interact to create a timbre perceptual space. The clustering models were employed to see how relationships between the signals might be created (based on known attributes of the signals themselves and parsing of the acoustic signal) and what features (those consistently found in the literature, as well as nonorthogonal relationships) specifically contribute to those groupings.

For the psychophysical experiment, we hypothesized (i) that the source waveforms consisting of a sawtooth wave with every second harmonic removed and the square with the twenty-third harmonic as the maximum (metameric pair) would be confused the most often (i.e., labeled as the ‘same’ though physically different), (ii) the waveforms falling at the extremes of the timbral continuum would be labeled correctly as ‘different’ the majority of the time (e.g., sawtooth wave comprised of 23 harmonics and square wave) and (iii) that the signals in the middle of the timbral continuum would be confused on occasion, but would be identified as ‘different’ a majority of the time. We additionally hypothesized that the confusion in item relationships in our crucial comparison (timbral metamers) would be preserved across vowel categories. Clustering/partitioning of the data was employed to see to what degree the physical structure and dimensions that can be recovered mathematically relate to behavioral and perceptual measures. Clustering of waveforms (and psychophysical responses) has been performed previously; we aim for a more rigorous approach using some of the data the auditory system evaluates (waveform dynamics, frequency and time-frequency transformations) to examine the

relationships between hypothesized and actual perceptual categories. *A priori*, we first hypothesized that there should be at least three clusters, one for each of the vowel categories used. Our second hypothesis was that the timbral metamers would be closest together in the clustered/partitioned space and this relationship would be constant across the different algorithms. We also entertained the possibility that the signals that were confused the most in the psychophysical data (i.e., the signals in the middle of the continuum) would be closest together within the clusters (aside from the metamers).

Part I: Psychophysical Evaluation of Signal Pairs

Materials and Methods

Participants

Thirty-two normal-hearing adult participants participated in signal comparison. Age range was 18-36 years, mean 20.9 years, median 20 years. Twenty-three participants identified as female, nine identified as male; one participant reported a non-English native language (Russian). Participants were compensated (\$7/half-hour) for their involvement. Three participants were removed from the final pool due to a prevalence of incorrect responses over all trials presented (> 50%). Presentation of stimuli and responses was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Before the start of each experiment, informed written consent was obtained from each participant. Participants were also asked to fill out a brief questionnaire documenting their musical or vocal training and experience. Eighteen participants

indicated that they had musical training, seven indicated that they were currently practicing their instrument. Additionally, five subjects indicated that they had some degree of ear training.

Stimuli

Experimental signals were generated with MATLAB (v7.8 R2009a), The Mathworks, Natick, MA). Four complex source signals were filtered using the transfer functions for the American English Vowels /a/, /i/, and /u/ (Hillenbrand et al., 1995). Vowel bandwidth measurements were likewise derived from ecological token measurements (Fant, 1972). Source signals were generated using Fourier synthesis and the transfer functions were generated from an all-pole filter derived from measurements of the mean frequency values (Hz) of the first three formants for male speakers. Signal duration was 250 msec (average approximate length of a vowel) with eleven msec \cos^2 onset and offset ramps. Signals were sampled at 44.1 kHz with 16-bit resolution.

The experimental materials were designed to be (i) ecologically valid, (ii) create timbral differences within and across signals classes and (iii) test the limits of the resolution of auditory perception (via timbral metamers). To that end, care was taken to select and synthesize source waveforms and filters (transfer functions) that satisfied these criteria. Since speech identification and processing is heavily reliant on timbre (Culling and Darwin, 1993; Fellowes et al., 1997; Gfeller et al., 1998; Dissard and Darwin, 2001), we chose to use as the basis for our experiments a source waveform that mimicked the glottal excitation pattern and modified its structure to

induce timbral differences. Further ecological validity was added by filtering the source waveforms through different vowel transfer functions as vowels are differentiated by their (i) timbre, (ii) vocal tract filtering characteristics and (iii) linguistic category. We chose vowel transfer functions that are well-separated timbrally and linguistically, namely the American English vowels /i/ (high front unrounded vowel), /u/ (high back rounded vowel) and /a/ (low back unrounded vowel). These vowels were chosen as they are the 'point vowels' and are well separated linguistically and perceptually. The source waveforms had a fundamental frequency (F0) of 150 Hz (gender-neutral fundamental voice frequency) and contained a maximum of twenty-three harmonics to approximate the glottal excitation pattern (Cleveland, 1977; Miller, 1989). Figure 1 illustrates the transfer functions for each of the signals used in the experiment, acting as filters on a white noise source signal to illustrate their properties. For each of the vowel types used in the experiment, the formant structure can clearly be observed; namely the peaks in the transfer function/vowel at each of the formant values, as well as the spacing between them.

The four experimental source signals consisted of a sawtooth waveform comprised of twenty-three harmonics, a sawtooth waveform with every fourth harmonic removed, a sawtooth waveform with every second harmonic removed, and a square wave where the maximum harmonic is the twenty-third; the sawtooth wave with every second harmonic removed and the square wave were the timbral metamers. Each source waveform was filtered digitally, for a total pool of twelve

signals. Signal power was equalized between signals (i.e., filtered source waveforms) via the root-mean-square (RMS) of the signal after filtering. Experimental materials for the psychophysical experiment were a subset of the signals clustered using different clustering methods and algorithms (see Part II below). Figure 2 illustrates the construction of the experimental materials, with the transfer function for the vowel /i/ acting as the filter and the sawtooth waveforms as the source waveforms. The left column illustrates the temporal structure of the waveform pre-filtering, the middle column, the effect of the filter (i.e., transfer function) on the structure of the waveform, while the last column illustrates the power spectral density estimate of the filtered waveform using Welch's method. The top row illustrates the sawtooth wave comprised of twenty-three harmonics, the middle the sawtooth waveform with every fourth harmonic removed and the bottom row the sawtooth waveform with every second harmonic removed. Figure 2 clearly illustrates the harmonic construction of the source signals in the temporal domain, the effect of the transfer function on the overall physical structure of the signal and the typical comb structure found in vowel and vowel-like sounds.

Experimental Procedure

Participants were randomly assigned to one of two separate experimental groups: one group was required to compare signals with the transfer functions corresponding to /a/ and /i/; the other /a/ and /u/. Participants were required only to indicate whether or not the two signals within a pair were the same or different, to provide a quick and efficient measure of participants' judgment and accuracy in

comparing pairs of signals, without them knowing *a priori* the basis on which the signals differed. Signal pairs were entirely within vowel categories (e.g., saw-23-a: sawtooth waveform, twenty-three harmonics, /a/ vowel; saw-4-rem-a: sawtooth waveform, every fourth harmonic removed, /a/ vowel) and counterbalanced; each pair was presented 10 times (20 counterbalanced pairs, 400 total trials). Participants were instructed to respond as quickly and accurately as possible. Participants were instructed not to chew chewing gum or other substances and to relax their jaw (to avoid pitch-shift effects) for the duration of the experiment (Hartmann, 1998). The inter-stimulus interval varied in the range between 272 and 578 msec and the inter-trial interval ranged between 714 and 1598 msec. Responses longer than 2000 msec were not recorded. Experimental signals and participant responses were collected via DMDX stimulus presentation software (Forster and Forster, 2003). Signals were presented using Dell Optiplex 320, running the Windows XP operating system with a SoundMax Integrated Digital HD Audio sound card. Stimuli were delivered to the participants binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). Participant responses were input using a Microsoft Sidewinder USB gamepad.

Data Analysis

Three participants who performed at chance or worse over all items (<50% correct over all items) were excluded from the final analyses; all three had been assigned to the /a/ - /u/ vowel transfer function pairing. For the remaining participants, *d'* values were calculated according to the independent observation

model and the differencing model of the 1AFC same-different (AX) paradigm to assess sensitivity in discriminating the signal pairs along with the observer criteria C , C' and $\ln\beta$. For the d' values, Hits were defined as when a participant responded with a correct response and the item pairs differed, and False Alarms when the response was incorrect and the items were physically identical, Misses were defined as when the response was incorrect and the items differed and Correct Rejections when the response was correct and the items identical. For participant sensitivity, we were additionally interested to see if musically/vocally trained participants showed greater discrimination sensitivity overall and signal discrimination within trials.

To compute d' for the independent observation model, we followed the method suggested by Macmillan and Creelman (1991). First, we estimated the proportion correct if item pairs were evaluated maximally:

$$Pc_{\max} = \text{cumnorm}([(z(\text{pH}) - z(\text{pF}))/2])$$

where Pc_{\max} is the maximum proportion correct, cumnorm is the cumulative normal function, z is the z -score, pH is the hit rate and pF is the false alarm rate. d' is then:

$$d' = 2z(0.5[1 + \text{sqrt}(2Pc_{\max} - 1)])$$

where z and Pc_{\max} are the same as before and sqrt is the square root function.

Observer criterion C was calculated as:

$$C = -[z(\text{pH}) + z(\text{pF})]/2$$

where z , pH , pF are the same as previously.

Criterion C' was calculated as:

$$C/d'$$

Criterion $ln\beta$ was calculated as:

$$ln\beta = [z(pF)^2 - z(pH)^2]/2$$

For the differencing model, we first calculated the differencing criterion k :

$$k = -z(pF/2)*sqrt(2)$$

where k is the criterion on which the signals differ. d' is given by finding the HIT rate that gives a specific d' value. A last analysis examined the correlation between the different observer criteria and the d' values via correlation using Spearman's method. Calculation of psychophysical parameters was performed using R 2.10.1 (R Development Core Team, 2010).

Participant response (proportion correct and reaction time (RT)) data were collected and analyzed using R 2.10.1. The data for each participant pool were analyzed by participant, vowel transfer function and item (trial). Statistical significance of the responses was assessed using General Linear mixed effects (GLM) models using the “languageR” statistical package (Baayen, 2010). Log (base10) RT and proportion correct were analyzed with Participant as a variable interaction; prior to execution of the final statistical models and analyses a list of possible factors was produced. *A priori*, we hypothesized the following factors would be most likely for the statistical analyses: Vowel, Vowel Height and Vowel Position as these are acoustic and linguistically relevant features of real vowel tokens, which are reflected in the construction of the synthetic signals employed in the experiment. The experimental data were analyzed first according to experimental assignment, and then on the pooled data (i.e., data from both experimental assignments). For each

experimental assignment the factors were Vowel (/a/ and /i/ or /a/ and /u/), Vowel Height (low or high), Vowel Position (front or back -- /a/ - /i/ assignment only) and Item. Analyses for the pooled data focused on Vowel Height, Vowel Position and Item. Interactions for the GLMs for each experimental assignment were as follows: (i) Vowel x Item, (ii) Vowel Height x Item, (iii) Vowel Position x Item and (iv) Item; any other potential factors (e.g., formant values or formant ratios) would have been collinear with the vowel and/or item categories. For the pooled data, Vowel, Vowel Height and Vowel Position were used as factors without any interactions.

Additional GLMs were implemented separately for participants by dividing the participant pool into two different groups, those that indicated they had musical training and those without. For the participants who indicated they had musical training, the factors were Training Type (formal or informal), Years playing the instrument, whether or not the instrument was currently being played, and whether or not the participant had ear training. The factor interactions for the analysis of the participants with musical training were (i) Training Type x Years played x Current playing x Ear Training, (ii) Training Type x Years played x Current playing, (iii) Training Type x Years played, (iv) Training Type, (v) interactions i-iv convolved with Item (to determine if the participants with musical training were more sensitive to the timbral differences). Two final statistical analyses compared the responses between those with musical training and those without (Crummer et al., 1994; Lakatos, 2000; Samson, 2003; Chartrand and Belin, 2006) and the effect of native language on sensitivity in discriminating within the vowel categories.

Results

Signal Discrimination Correctness

Proportion correct for each experimental group was calculated for each item pair presented and overall for each transfer function. Contrary to our hypotheses, we observed a surprising asymmetry in percent correct identification in both experimental assignments; Figure 3 shows the proportion correct over all items for each vowel in the pooled experimental data. For both experimental assignments, signals in the /a/ category were discriminated correctly much less than the signals in the /i/ or /u/ categories. For the /a/ signals, the percent correct was ~50%; for the /i/ and /u/ signals, the percent correct was ~80%.

Figures 4 and 5 plot the proportion correct for each item pair; Figure 4 for the /a/ - /i/ experimental assignment and Figure 5 for the /a/ - /u/ experimental assignment. In accordance with our hypotheses, the signals with perceptually identical source waveforms (sawtooth with every second harmonic removed and the square wave) were identified as being the same, when physically different; this pattern held across all vowel categories. For the /i/ signal pairs, no signal pair had a percent correct identification less than 76%; for the /u/ signal pairs, no pair had a percent correct identification less than 42%. Interestingly, for both the /i/ and /u/ signals the non-metameric pairs with the lowest percent correct were the ones where a sawtooth wave comprised of 23 harmonics was paired with a sawtooth wave with every fourth harmonic removed.

The asymmetry observed between the different vowel categories was due to signals in the middle of the timbral continuum for the /a/ signal class. Aside from the /a/ signal pairs where each signal in the pair was identical, which had percent correct values of ~90%, the signal pairs a-23-4, a-23-2, a-23-square, a-4-2 and a-4-square had percent correct values of ~10-45% for both experimental assignments. These observations suggest that the formants and/or formant ratios differentially affect the filtering of the source signals and possibly the shape of perceptual space of the signals and their relationship to one another (see *Discussion*).

Since reaction times were not statistically different between items, we decided to focus solely on the responses to the items; namely (i) what factors were most likely to predict a correct or an incorrect response and (ii) what factors predicted the asymmetry seen between the different vowels. Due to the rather large degrees of freedom values for our initial statistical analyses, we decided not to evaluate each individual trial response, but to collapse the data by taking the mean response for each factor. We focused on the proportion correct for the factors Vowel, Item, and SCG differences for the individual experimental assignment data and Vowel, Item, SCG, and Vowel Height and Vowel Position for the pooled data. Collapsing the data resulted in a more conservative analysis with respect to the responses given on each trial. The data in the resulting data frames were then evaluated using GLMs.

For the /a/ - /i/ experimental assignment, analysis of the collapsed data revealed a significant difference in the proportion correct between vowel categories, with the /i/ signals as a predictor of a correct response (coef = 0.284, SE = 0.027, t = 10.560). For the analysis of proportion correct by item, significant predictors were observed to be all /a/ trials where the signals were not identical (a-2-square: coef = -0.831, SE = 0.058, t = -14.266; a23-2: coef = -0.481, SE = 0.058, t = -8.259; a-23-4: coef = -0.726, SE = 0.058, t = -12.979; a-23-square: coef = -0.463, SE = 0.058, t = -7.937; a-4-2: coef = -0.544, SE = 0.058, t = -9.332; a-4-square: coef = -0.591, SE = 0.058, t = -10.136), the /i/ metamer trials (coef = -0.697, SE = 0.058, t = -11.960) and a single /i/ signal comparison (i-23-4: coef = -0.116, SE = 0.058, t = -1.984). The results of the statistical analyses indicate that these trials were predictors of an *incorrect* response.

For the /a/-/u/ assignment, statistical analysis revealed the response asymmetry in the proportion correct between the /a/ and /u/ signals was significant, with the /u/ signals as a predictor of a correct response (coef = 0.232, SE = 0.024, t = 9.704). As with the /a/ - /i/ experimental assignment, all the /a/ trials where the signals were not identical were significant predictors of an *incorrect* response (a-2-square: coef = -0.877, SE = 6.492e-02, t = -13.507; a23-2: coef = -0.488, SE = 6.492e-02, t = -7.524; a-23-4: coef = -0.788, SE = 6.492e-02, t = -12.145; a-23-square: coef = -0.492, SE = 6.492e-02, t = -7.583; a-4-2: coef = -0.604, SE = 6.492e-02, t = -9.301; a-4-square: coef = -0.615, SE = 6.492e-02, t = -9.479) as well as the /u/ metamers (coef = -0.819, SE = 6.492e-02, t = -12.619).

For the pooled experimental data, the proportion correct responses to the /i/ (coef = 0.279, SE = 0.023, t = 12.050) and /u/ (coef = 0.238, SE = 0.025, t = 9.390) signals were found to significantly differ from those of the /a/ signals, being predictors of an overall correct response. For the effect of vowel height, the low vowel signals (/a/ signals) were found to be predictors of an incorrect response (coef = -0.261, SE = 0.018, t = -13.970). For the effect of vowel position the front vowel (/i/ signals) overall (coef = 0.256, SE = 0.026, t = 9.871) were found to be predictors of a correct response. For the last two sets of statistical analyses (effects of musical training and native language), there were no differences in response patterns between participants with and without musical training as well as between the native and non-native English speakers. There was no difference in signal evaluation for musically trained participants with and without ear training.

Based on the asymmetries observed in the response data for the /a/ signals as well as the surprising findings from the signals in the middle of the continuum for the /a/, /i/ and /u/ signals, we conducted *post hoc* tests on (i) the /a/ signals and (ii) differences in the spectral center of gravity (SCG) between the signals compared within a trial as a factor. We employed a modified version of the formula for SCG, derived from Caclin (2005). The spectral center of gravity measurement (which involves amplitude-weighted spectral information) was derived via power spectral density (PSD) estimates of the digitized, RMS-equalized waveforms presented to the experimental participants using Welch's method implemented in MATLAB's signal

processing toolbox. The PSD was estimated only for the present in the spectrum (harmonics and spectral envelope). SCG was computed as follows:

$$SCG = (\sum_n n \times A_n) / \sum_n A_n$$

where n denotes the frequency in Hz, and A is the amplitude in decibels (dB). We also included a version of the SCG where frequency is log-transformed (log base 10) to improve ecological validity; a log-log representation of the signals would more accurately represent how the auditory system parses the signal (Hartmann, 1998). For the GLMS implemented, the SCG measurements were the only factors considered. The *post hoc* /a/ GLMs only considered item (trial) and the SCG measures as the only factors considered.

Statistical evaluation of the SCG measures (linear and log frequency) found significant predictors to be trials that predicted both correct and incorrect responses. For the /a/ - /i/ experimental assignment, the trials where the difference in SCG (linear frequency) were predictors were the metamer trials (a-2-square: coef = -0.863, SE = 0.063, $t = -13.620$; i-2-square: coef = -0.728, SE = 0.063, $t = -11.498$) as well as the /a/ trials where the signals within the trials were not identical (a-23-square: coef = -0.494, SE = 0.063, $t = -7.970$; a-23-4: coef = -0.788, SE = 0.063, $t = -12.435$; a-23-2: coef = -0.5125, SE = 0.063, $t = -8.093$; a-4-square: coef = -0.622, SE = 0.063, $t = -9.820$; a-4-2: coef = -0.575, SE = 0.063, $t = -9.080$). The only other /i/ signal comparison found to be a predictor, were the trials that had the lowest proportion

correct (i-23-4: coef = -0.147, SE = 0.063, t = -2.319). For the logarithmically weighted SCG measurements, the metamer trials were found to be predictors of an incorrect response (a-2-square: coef = -0.716, SE = 0.063, t = -11.300; i-2-square: coef = -0.581, SE = 0.063, t = -9.178), as well as the /a/ trials where the signals differed (a-4-2: coef = -0.428, SE = 0.063, t = -6.760; a-4-square: coef = -0.475, SE = 0.063, t = -7.501; a-23-4: coef = -0.641, SE = 0.063, t = -10.116; a-23-2: coef = -0.366, SE = 0.063, t = -5.774; a23-sq: coef = -0.347, SE = 0.063, t = -5.477). These observations parallel the findings from the previous statistical analyses. Additionally, several of the /i/ trials were found to be predictors of a correct response (i-23-2: coef = 0.138, SE = 0.063, t = 2.171; i-23-sq: coef = 0.147, SE = 0.063, t = 2.319). The last predictor of a correct response was having a log frequency SCG difference of zero (coef = 0.139, SE = 0.063, t = 2.190). Interestingly, the log frequency SCG values and the difference between the signals evaluated within a trial based on the log frequency SCG values seemed to be a better indicator of signal relatedness at least for the metamers; the resulting values indicated the signals more closely resembled each other as opposed to the linear frequency SCG measures (/a/ metamers: 6.46e-05; /i/ metamers: 8.77e-05).

For the /a/ - /u/ experimental assignment, both SCG measures found significant predictors of correct and incorrect responses. For the linear frequency SCG measure, predictors of an incorrect response were the timbral metamers (a-2-square: coef = -0.385, SE = 0.651, t = -5.907; u-2-square: coef = -0.327, SE = 0.651, t = -1.890) as well as several, but not all, of the /a/ signal trials where the signals

within a trial differed (a-23-4: coef = -0.297, SE = 0.651, t = -4.548; a-4-sq: coef = -0.123, SE = 0.651, t = -1.890). Predictors of a correct response were the trials where the signals were identical (coef = 0.475, SE = 0.651, t = 7.302) as well as several of the /u/ trials where the signals differed (u-23-square: coef = 0.492, SE = 0.651, t = 7.561; u-23-2: coef = 0.442, SE = 0.651, t = 6.852; u-4-square: coef = 0.442, SE = 0.651, t = 6.793; u-4-2: coef = 0.419, SE = 0.651, t = 6.438). For the log frequency SCG measures, the timbral metamers were found to be predictors of an incorrect response (a-2-square: coef = -0.860, SE = 0.065, t = -13.209; u-2-square: coef = -0.802, SE = 0.065, t = -12.323), as well as the /a/ signals where the signals within a trial differed (a-4-2: coef = -0.587, SE = 0.065, t = -9.015; a-4-square: coef = -0.599, SE = 0.065, t = -9.192; a-23-4: coef = -0.772, SE = 0.065, t = -11.850; a-23-2: coef = -0.472, SE = 0.065, t = -7.243; a-23-square: coef = -0.475, SE = 0.065, t = -7.303). A single /u/ trial was found to be a predictor an incorrect response (u-23-4: coef = -0.479, SE = 0.651, t = -7.361). No predictors for correct responses were observed.

The *post hoc* analyses for the /a/ trials focused only on item as factor; responses were collapsed as with the previous statistical analyses. The *post hoc* analyses confirmed the results of the experimental assignment analyses, namely that the /a/ trials where the signals differed were predictors of an *incorrect* response (a-2-square: coef = -0.852, SE = 0.046, t = -18.344; a-23-2: coef = -0.484, SE = 0.046, t = -10.435; a-23-4: coef = -0.771, SE = 0.046, t = -10.249; a-4-2: coef = -12.291, SE = 0.046, t = -0.571; coef = -12.960, SE = 0.046, t = -0.602).

d' (Participant Acuity) Values

Figure 6 displays the histograms of the d' values for both the independent observation and differencing models of signal discrimination. *A priori*, it could be hypothesized that the differencing model was the more likely descriptor of participant behavior, due to the large number of items employed (Macmillan and Creelman, 1991). The values of d' for each participant according to the independent observation model ranged from 0.535 to 3.379, with the majority of participants having d' scores in the range from 1 to 3. For the differencing model, d' ranged from 0.622 to 4.180, with the majority of participants exhibiting d' scores in the range from 1 to 4. The observed d' values are consistent with the proportion correct values observed for each participant and the difficulty of the task. While the timbral metamers would have decreased the overall participant sensitivity (perhaps only slightly), the signals in the middle of the continuum (especially the /a/ signals) seem to have kept the observer sensitivity from being greater. Examination of the histograms for d' scores indicated that participants overall did well in discriminating the signals.

Figure 7 displays histograms of the various observer criteria for the independent observation model (C , C' and $\ln \beta$) and the differencing model (k). Observer criteria C values for the independent observation model ranged from 0.092 to 1.138, with the majority of values lying in the range from 0.3 to 1.2. Observer criteria C' values ranged from 0.064 to 0.756, with most values lying in the range from 0.3 to 0.5. Observer criteria $\ln \beta$ values ranged from 0.020 to 3.103, with the majority of values lying in the range from 0.4 to 1.2. Criteria k for the

differencing model ranged from 1.175 to 3.867, with the majority of values lying in the range from 2 to 3.2. For the independent observer model, weak or nonexistent correlations were observed for the criteria C (0.486) and C' (-0.122), while a strong correlation was observed between the criteria $\ln\beta$ (0.808) and d' scores. For the differencing model, a strong correlation (0.794) was observed between differencing criteria k and d' scores.

Discussion

The psychophysical measurement of the ecologically approximate signals in the experiment presented has demonstrated some of the complex interactions between the structure of a source signal and its spectral filter. While our main hypothesis concerning the timbral metamers was supported by the data, we observed surprising findings regarding the /a/ signals employed, and to some extent, some of the /i/ and /u/ signals also.

Concerning the timbral metamers, these source signals were engineered to have identical spectral structures. A more in-depth examination, with respect to the calculation of the PSD of the metamers, indicates that even after RMS equalization for power, there was a slight difference in the amplitudes of the sawtooth wave with every second harmonic removed and the square wave source signals. However, this amplitude (and hence perhaps loudness difference) did not seem to affect the comparison of these signals across all synthesized vowel classes; the timbral metamers were confused as being the same, in spite of the minimal phase difference.

This is in line with the previous literature that phase plays a minimal effect regarding the timbre perception of steady-state signals. Our data also agree with the timbral literature in that the timbre of steady-state signals is completely determined by spectral structure, a duplex perception where the spectral envelope (formant structure) and the source signal (harmonics/frequencies present) combine to determine the timbre.

Regarding the unique findings pertaining to the /a/ signals and the /i/ and /u/ signals in the middle of the timbral continuum, we hypothesize that this is due to the filtering characteristics of the spectral envelope (formant structure), especially for the /a/ signals. The transfer functions of each of the synthesized vowels serves to attenuate and/or boost specific frequencies. Based on the pattern of attenuation/boosting, each signal will have a specific center of gravity and have different perceptual distances from each other within a perceptual category. While categories need not be mutually exclusive, especially when linguistic experience causes difficulty in the processing of specific sound categories, it is still important to learn in a general sense within and across categories (Pastore et al., 1990). Evidence from this is based on the statistical analyses where differences in SCG was a factor in GLMs. The majority of signals that were confused the most often had very small differences in spectral center of gravity, indicating that this was perhaps a very salient cue when comparing signals within a trial. Additionally, because the signals employed in the experiment were based on vowel data, the formant structure may have also induced the asymmetry observed. For example, vowel identity seems to be

highly correlated with the F1/F2 ratio and the formant values may be ‘averaged’ in a sense. The closeness of the formants, then, may result in an inability to distinguish signals within a particular class/category. The psychophysical results also support the hypothesis and agree with previous findings regarding SCG as a dimension or factor that is used in signal discrimination. More specifically, the SCG values with log frequency maybe a better predictor. This is borne out by the statistical analyses; the signals with very small differences in log frequency SCG were more often predictors of an *incorrect* response with few exceptions. Those exceptions that did occur may be a result of the effects spectral envelope structure influencing factors such as virtual pitch.

With the psychophysical results in mind then, the clustering/partitioning of the waveforms and their various representations can potentially aid in the understanding of the ordering and creation of functional classes/groupings and the relationships of signals within classes. They can also help to tease apart to what extent the implementation of clustering algorithms on the various representations (especially biologically motivated representations) match the psychophysical responses observed.

Part II: Clustering of ecologically-based synthesized signals

Signals

Experimental signals were generated with MATLAB (v7.8 (R2009a), The Mathworks, Natick, MA). Complex source signals were filtered using the transfer functions for the American English Vowels /a/, /i/, and /u/ (Hillenbrand et al., 1995).

Vowel bandwidth measurements were likewise derived from ecological token measurements (Fant, 1972). Source signals were generated using Fourier synthesis and the transfer functions were generated from an all-pole filter derived from measurements of the mean frequency values (Hz) of the first three formants for male voices. Signal duration was 250 msec (approximate average length of a vowel) with eleven msec \cos^2 onset and offset ramps. The experimental materials were designed in the same manner as the signals employed in the psychophysical experiment, but with a larger signal set. Seven different source waveforms were employed to create a timbral continuum. The signals were as follows: sawtooth waveform approximating the glottal excitation pattern (twenty-harmonics in spectrum), sawtooth waveforms with every sixth, fifth, fourth, third, and second harmonic removed and a square wave where the maximum harmonic is the twenty-third. The source signals had the same F_0 , were filtered using the same transfer functions, the same onset and offset ramps as in the psychophysical experiment for a total of twenty-one signals. Signals were equated for power via RMS after filtering.

The signals were represented in three different ways: (i) the physical structure of the waveform, (ii) the PSD of the waveform and (iii) a wavelet-based time-frequency cochlear-model transformation of the physical structure. All signal manipulations were performed in MATLAB. Implementation of the cochlear representation required resampling to 16 kHz. The physical structure (temporal evolution) of the signals was obtained after signal creation via MATLAB. The PSD estimate via Welch's method of the signals was computed using MATLAB's Signal

Processing Toolbox. The wavelet-based cochlear model representations were performed using the NSL Toolbox (Chi and Shamma) in MATLAB.

Clustering Methods

The signal representations were clustered/partitioned using a variety of algorithms and methods: (i) k-means clustering, (ii) hierarchical agglomerative clustering and (iii) image classification via image similarity measures in conjunction with MDS. K-means and hierarchical clustering as well as the statistical assessment of the classifications, were performed using R 2.10.1. Calculation of image similarity was performed using MATLAB, while MDS was implemented using R 2.10.1.

K-means clustering was implemented using the default k-means clustering algorithm, with aggregation based on mean values. A more robust version of k-means clustering was also implemented using the “fpc” package in R (Hennig, 2010), with aggregation based on mediods instead of means. Clustering was initially implemented with three clusters extracted (corresponding to each of the vowel categories); after performing this initial clustering, we used calculation of the within group sum of squares (ordinary k-means) and the optimum average silhouette width (“fpc” package, robust k-means) to determine the optimum number of clusters.

Hierarchical agglomerative clustering was implemented using Ward’s method to create the clusters and Euclidean distance to calculate the distance matrix. The resulting dendrogram was initially cut into three clusters in accordance with our

initial hypothesis; additional dendrogram cutting was also implemented based on the secondary hypotheses regarding linguistic information and signal structure, as well as the optimum clusters found via k-means clustering. Statistical significance of the resulting clusters was evaluated using, the “pvclust” package in R (Suzuki and Shimodaira, 2009), which evaluates the dendrogram results via multiscale bootstrap resampling (1000 bootstrap resamplings).

The cochlear transformations of the signals resulted in a three-dimensional representation: time x frequency x intensity. The parameters for the cochlear representation were a frame jump of 4 msec, a time constant of 16 msec and a nonlinear factor of -1. The final dimensions were 32 channels by 128 time points. The resulting spectro-temporal profiles were compared as images, where time and frequency replace the spatial axes and luminance is replaced by intensity. This may have some biological relevance, especially at the cortical level, as texture analysis may be a common method of evaluating signals (Yau et al., 2009; Bruckert et al., 2010; Overath et al., 2010). Figure 8 illustrates the cochlear model representation of the sawtooth waveform approximating the glottal excitation pattern for the vowel /a/. The cochlear representation emphasizes the concentration of energy in the F1-F2 formant region as well as the harmonic structure of the signal.

To calculate the spectro-temporal profile of the cochlear and cortical representations of the signals, we used the structural similarity index (SSIM) for image quality assessment (Wang et al., 2004). Two distinct spatial filters were used:

the first was a symmetric Gaussian (window: 11 x 11 square matrix; standard deviation: 1.5) and the second was a contrast enhancement filter (window: negative of Laplacian, 3 x 3 matrix; alpha: 0.2). The Gaussian was employed as a smoothing filter to compare gross aspects of the signal representations (averaging within image followed by inter-image comparison) and the negative of Laplacian filter to compare fine structure features (differences/contrasts within image followed by inter-image comparison). Figure 9 illustrates the effects of the Gaussian smoothing filter and the contrast enhancement filter on the experimental signals. The effect of the Gaussian in Figure 9a is to highlight the gross features of the signal; the energy concentrated in the F1-F2 range as well as the third formant is clearly visible, while contributions from the harmonic structure of the source waveforms is minimized. The effect of the contrast enhancement (Figure 9b) is to emphasize the harmonic structure of the source waveform (horizontal bands), including and beyond the first formant.

After applying the spatial filter and calculating image similarity, symmetric matrices were then computed, with the indices from the spatial smoothing and comparison interpreted as similarity measures (values of the symmetric matrix produced were subtracted from 1 to produce dissimilarity values) and the indices from image contrast enhancement and comparisons interpreted as dissimilarity measures (no subtraction was needed). The symmetric matrices were then evaluated using metric and Kruskal's nonmetric multidimensional scaling in R; nonmetric multidimensional scaling was implemented using the "MASS" package (Shepard, 1962; Kruskal, 1964a, b; Venables and Ripley, 2002). MDS dimensional reduction

solutions were calculated in two and three dimensions (Plomp et al., 1967; McAdams et al., 1995; Sinnott et al., 1997; Lakatos, 2000; Caclin et al., 2005). K-means and hierarchical clustering of the cochlear signals was accomplished by reshaping each of the transformations into one-dimensional vectors.

Results

Out of the clustering/partitioning methods employed, hierarchical clustering and MDS were most effective in clustering the signals. Results from k-means clusters were not very parsimonious, even though the first two principal components explained ~80% of the variance. This was a function of the manner in which the data were apportioned and represented; graphical representation of the clusters indicated that the data were clustered as to specific time point or frequency in Hz, instead of according to signal structure itself. This issue was not present with hierarchical clustering and MDS of the signals.

Hierarchical Clustering Signal Division

Figure 10 illustrates the dendrograms for the different representations of the signals with three clusters, in line with our initial hypothesis. Contrary to our hypotheses, the signals were not clustered completely according to their linguistic/phonetic categories. While this is unsurprising in one sense, namely that there is little information present in the signals (aside from the structure imposed by the formant values and periodicity) that would yield this potential separation of the signals, upon initial inspection there does not seem to be any consistent relationship between the dendrogram branches.

The clustering of the temporal waveforms with the initially specified three cluster solution seems to be based partially on vowel/formant information and partially on the source waveform. This is evidenced by almost all the /a/ signals being in one cluster (low vowel), the /i/ and /u/ signals being in another (high vowels) and the waveforms with square wave source signals in a separate cluster. For the PSD, the link between clusters and vowel/formant information was somewhat stronger; with a few exceptions, the signals were clustered within each vowel (/a/, /i/ or /u/). Hierarchical clustering of the cochlear representations was not able to be implemented; this was due to errors in the calculation of the distance matrix. Calculation of the distances between vectors resulted in NaN values due to the presence of negative similarity and dissimilarity values.

Analysis of the within sum of squares indicated that 4-5 clusters was ideal for the waveforms and 4-6 clusters for the PSD of the signals. Initial observation of the four cluster solutions for the temporal waveform indicated that this particular solution was better at dividing the signals according to linguistic categories, though some signals (a-5-rem, i-2-rem, u-3-rem) were not clustered with their vowel categories. The signals with square wave source waveforms were still placed in separate clusters, as in the three cluster solution. The five cluster solution for the temporal waveforms added a separate cluster for a single signal (sq-a-23), while the other clusters were unchanged.

As with the three cluster solution for the PSD of the signals, the four cluster solution for the PSD of the signals primarily grouped the data by vowel information, with the exceptions being the timbral metamers for the /i/ signals (i-2-rem,sq-i-23) and a separate cluster for one signal (a-5-rem). The five cluster solution added a cluster with the metamers for the high vowels. The six cluster solution was identical to the five cluster solution, with the addition of a cluster for the /a/ metamers.

Figure 11 displays the dendrograms for the bootstrap supported clusters. Multiscale bootstrap resampling of the dendrograms is interpreted as follows: the numbers in red are approximately unbiased p-values, while the numbers in green are the bootstrap probability values. Rectangles surround those clusters with > 95% confidence. The number of bootstrap-supported clusters was less than that of the suggested optimal number of clusters from the within sum of squares analysis. For the temporal waveforms and PSD, there were only two clusters supported after multiscale bootstrap resampling. For the temporal waveform, the clusters seemed to be grouped by vowel information (most of the high vowel signals were grouped together), while the /a/ signals and square wave source waveform signals were in the second cluster. For the PSD, there were two bootstrap-supported clusters; the clusters grouped the high vowel signals (/i/ and /u/) separately from the low vowel signals (/a/).

A last examination of the dendrograms and their branches examined how closely signals were clustered together, based on SCG (linear and logarithmic

frequency scaling) as a complexity measure. The raw waveform dendrograms and branches seem to have no relationship to the SCG values, as they vary widely within clusters. The PSD clusters seemed to have a greater relation to the SCG values for the three cluster solution (based on the ranges of SCG values of the signals within the cluster) and for the bootstrap supported clusters. The /i/ and /u/ signal SCG values ranged from 1933.447 to 2034.875 (linear frequency) and 3.105 to 3.135 (log frequency); for the /a/ signals, the SCG values ranged from 2075.842 to 2094.353 (linear frequency) and 3.135 to 3.149 (log frequency). This implies that the linear SCG (or central tendency) may have been a dimension that the algorithm was able to extract, since the different signals classes had well-separated ranges.

MDS Signal Division

Implementation of MDS performed on the cochlear representations of the signals was successful except for non-metric MDS on the contrast-enhanced (fine structure comparison) two- and three-dimensional representations; this was due to a zero/negative distance between two of the objects. For the classical MDS implementation, the constant c^* was added in order to prevent negative eigenvalues from occurring. Non-metric MDS yielded no easily interpretable results for the gross structure comparison. There were two overall groups/clusters of signals from the non-metric MDS implementation; within these groups however, the signals had essentially identical coordinate values and this solution was uninformative.

Figures 12-14 illustrate classical MDS solutions for the gross feature and fine structure enhanced cochlear transformations. Significant areas are shaded to

highlight their position in the MDS space. Figure 12a plots the two-dimensional classical MDS solution for the gross feature comparison, with each signal identified by its vowel category and source waveform harmonic structure. The gross feature comparisons ($\lambda_{1,2} = [1.390, 0.203]$, GOF = 75%) produced a better fit to the data than the fine structure comparisons ($\lambda_{1,2} = [197.60, 169.015]$, GOF = 16%) for the two-dimensional solution. From this visualization of the signals and their relationship to each other, it appears that the first principal coordinate separates the signal according to vowel height (high vowels are on the right of the abscissa); the high vowel signals (/i/ and /u/) are distinctly separated from the /a/ signals along this coordinate. With regards to vowel category, it does not appear that the second coordinate groups or indexes vowel category in any meaningful manner. Rather, the second principal coordinate seems to separate the high vowel metamers from the rest of the signals, as with multiscale bootstrap resampling of the PSD. The signals in the lower right corner of the figure are the high vowel metamers. In line with our hypotheses, it is not surprising the metamers would be located closest to one another. However, it is somewhat surprising that the coordinates for the /i/ and /u/ metamers do not overlap, at least not completely. Inspection of the /a/ metamers (located at (-0.20,-0.36)) indicates that these signals almost completely overlap, in line with our initial hypotheses. The difference between the high vowel and low vowel metamers may be a result of differences in spectral envelope as well as the phase difference (though slight) between the metamer source signals. Further examination of the two-dimensional gross feature comparison indicates that the /u/ and /i/ signals exhibited a high degree of self-similarity and were tightly clustered within their vowel class, with

a single exception (u-3-rem). Aside from the metamers, the /a/ signals were more spread out in their distribution in the principal coordinate space, with one signal (a-5-rem) being considerably further from the rest of the /a/ signals. Interestingly, this is the signal that was furthest away from the other /a/ signals when hierarchical clustering was implemented. The data visualization indicates that harmonic structure of the source waveforms was not indexed by either principal coordinate. Analysis of the two-dimensional solution space according to linear and log frequency SCG values indicated that the density of the clusters seemed to be partly based on log frequency SCG values.

Figure 12b illustrates the two-dimensional classical MDS solution to for the fine structure comparison. Unlike the two-dimensional solution for the gross feature comparison, there is no clear division within either of the principal coordinates of the signals according to vowel category. The /a/ signal with every fifth harmonic removed was still furthest from the rest of the signals, as in the hierarchical clustering and gross feature comparisons clusterings.

Figures 13a and 14 illustrate the three-dimensional MDS solutions for the gross signal feature comparisons. The three-dimensional representations exhibited better GOF values than the two-dimensional solutions. As with the two-dimensional solutions, the gross feature comparisons fit the data better ($\lambda_{1,2,3} = [1.390, 0.203, 0.182]$, GOF = 83%) than the fine structure comparison ($\lambda_{1,2,3} = [197.160, 169.015, 154.004]$, GOF = 23%). The figures provide visualization in two ways: (i) three-

dimensional plots of the MDS signal space and (ii) two-dimensional plots where each of three coordinates are plotted against one another. As in the two-dimensional solution, the three-dimensional classical MDS solution for the gross feature comparison seemed to be better at classifying the data according to vowel category.

Figure 13a highlights the major divisions of the three-dimensional MDS space for the gross feature comparison. There are three major groupings. First, the /a/ signals (red ellipse) are separate from the /i/ (gray rounded rectangle) and /u/ signals (blue rectangle) along Coordinate 1. Second, the /i/ and /u/ metamers are separate from the remaining /i/ and /u/ signals along Coordinate 2. Lastly, the third coordinate seems to separate the /u/ signals from the /i/ signal (black line). Figure 13b illustrates the three-dimensional MDS space for the fine structure comparisons. As with the two-dimensional solution, the fine structure space is not highly ordered.

Examining the MDS space for coordinates 1 and 3 indicates that the third coordinate seems to separate the /i/ and /u/ signals from one another. The third dimension may also index what might be interpreted as tongue position; the /i/ signals are located 'higher' than the /u/ signals. Examination of the MDS space for coordinates 2 and 3 seems to indicate that not only are the metamers separated, but an interaction between the points in the space seems to index where in timbral continuum the signals lie. For example, the back vowel (/a/ and /u/) signals that are not metamers lie mostly between the points (-0.05, 0.05) for coordinate 2 and the points (-0.15, 0.1) for coordinate 3. Figure 14 illustrates the 3D gross feature

comparison by plotting each coordinate against each other. The relationships in terms of vowel category mentioned above are much more easily identifiable when visualized in this manner. Fine structure spatial configuration did not seem to index the signals in any meaningful way; this resulted in their not yielding an easily interpretable representation of the space as previously.

Based on the results of the fine structure comparison analysis, we conducted *post hoc* tests examining whether or not the fine structure comparisons were better at evaluating relationships within vowel categories, as opposed to the gross feature comparisons, which seemed to index mostly vowel categories along with some indexing of within-category differences.

The two-dimensional MDS solutions for the within-vowel signal comparisons produced better GOF values than when the total pool of signals was considered together, with the high vowel signals exhibiting slightly better GOF values (/ɑ/: $\lambda_{1,2} = [3.139, 2.877]$ GOF = 40%; /u/: $\lambda_{1,2} = [22.846, 13.405]$, GOF = 46%; /i/: $\lambda_{1,2} = [81.387, 54.072]$, GOF 46%). Overall however, none of the dimensional reduction fits was of high quality, with all the GOF values <50%. Within the MDS signal space for each vowel, there does not seem to be a high degree of order, and for only one vowel category (/ɑ/), the metamers are positioned close to one another in the space. This is unsurprising in one sense, as all the GOF of fit values < 50%, but could also be a result of using having only seven source signals for each vowel, resulting in a sparse space. Neither of the principal coordinates seems to index signal structure in

any systematic way. The three-dimensional solutions for the within-vowel comparisons fit the data much better than the two-dimensional solutions; an asymmetry between the high and low vowel signals was still observed however: (/a/: $\lambda_{1,2,3} = [3.139, 2.877, 2.698]$, GOF = 58%; /u/: $\lambda_{1,2,3} = [22.846, 13.405, 12.937]$, GOF = 62%; /i/ = $[81.387, 54.072, 53.179]$, GOF = 64%).

Discussion

The clustering algorithms, especially the texture analyses, were successful in dividing the various signals according to their basic physical and sensory features. The observation that the signals could be grouped according to linguistic information/category was somewhat surprising, though this grouping may be a result of some of the characteristics of the signal categories explored in the experiment (see below). For the textural analysis of a possible signal space, the dimensions the signal were clustered/portioned along seemed to reflect the most salient attributes of steady-state signals reported in timbre psychophysical studies (spectral envelope, harmonic content/structure, spectral center of gravity/centroid, source waveform). Somewhat unsurprisingly, the gross feature identification of the texture analysis of the signals seemed to correlate well with the most salient features of the signals and their overall categories, while the fine structure comparisons were better at mapping a signal space within categories.

The successful implementation of various clustering algorithms and signal space measures using different signal representations demonstrates the feasibility of

grouping signals according to their basic features. This implies that (especially with regards to the cochlear texture analysis) absent some sort of context (other phonetic cues, temporal and spectral evolution of a sequence) auditory signals can be grouped and characterized by essentially only their basic sensory features. While in a psychophysical context this is not surprising, as certain features are more salient than others, it is surprising from a purely mathematical standpoint. There was no particular reason *a priori* for the signals to cluster according to their structure for the hierarchical clustering algorithms (even though we hypothesized this might be the case), but the implementation was successful, with the bootstrap supported clusters in some cases largely grouping the signals according to phonetic/linguistic information. A possible explanation for the results observed is that some vowel characteristics were able to be extracted by the algorithms; vowels are distinguished (especially in the frequency domain) by their formant spacing and spectral peaks and valleys. (Stevens, 1998)

While there is evidence that it is possible (and likely) that textural analysis of sensory signals is a common method of evaluating information in different modalities in cortex, this has not been shown in the sensory periphery. The cochlea is well-studied and performs a time-frequency-intensity analysis of auditory signals, lending itself to a texture analysis of the signals. The results of MDS on the gross features and fine structure comparisons of the cochlear representation textures suggest that textural analysis may well be a viable way of signal evaluation at the periphery. The results of MDS on the different textures additionally indicate that emphasis of

different features work better at different scales: the gross feature comparison were better at separating the signals based on vowel category and source waveform structure, while the fine structure comparisons were better at within-category separation (though not with very high GOF values). The gross features MDS space may have reflected those attributes that are more salient to the observer/algorithm and agree quite well with previous psychophysical data (e.g. spectral envelope structure, source waveform structure, SCG). The limits of the fine structure comparisons may be due to the relatively low number of signals (yielding a low-density space) used or perhaps the characteristics of the window or spatial filter employed.

While it is tempting to say the algorithms employed are capable of extracting phonetic data from the signals used in the cluster analyses, the results (aside from the biologically-based cochlear transformations) may not entirely support this view. The spectral envelopes of the signals were taken from human vowel data, and as such, this may have influenced the groupings of the signals. Certain vowels have certain characteristics (e.g., unique spectral tilts) and these features have been known and studied since the 19th century by linguists and auditory scientists. It is almost certainly the case, even though the signals used are completely synthetic and lack certain features of real vowel tokens (e.g., formant transitions breathiness), that these characteristics were reflected in the signals employed. It is thus unclear as to what extent phonetic information contributes to groupings of the signals (though phonetic information seems to be represented in some manner).

That being said, the signal representations (especially the cochlear textural representations), seem to lend themselves to clustering/grouping by other algorithms. For the MDS signal spaces, their orientations could possibly be learned by neural network algorithms (e.g., a simple perceptron or support vector machines). Networks and models incorporating biological parameters or that are biologically inspired could be used to group sensory signals into a signal space. Additionally, the use of various clustering algorithms and signal representations could be applied to such problems as learning signal generalities across speakers, differences in signal duration, differences in presentation, etc.

Overall, we find it is possible to examine the physical and perceptual relationships of auditory signals using a variety of mathematical and statistical techniques in addition to using biologically-motivated signal representations. The results of the analyses presented demonstrate that to some extent, low-level signal attributes and peripheral mechanisms correlate well with abstract mathematical transformations and perceptual saliency.

General Discussion

Our initial hypotheses for the psychophysics were that the timbral metamers would be confused the most, while the other signals would be correctly discriminated the majority of the time. We based this hypothesis on the steady-state nature of the signals; differences in spectral center of gravity and spectral content would most likely determine how the signals were discriminated. For the clusters, working from

the basis of spectral envelope information, we hypothesized the signals would be grouped according to vowel category while making no predictions about the orientation of the space.

While our hypothesis concerning the timbral metamers was preserved across categories in the psychophysical experiment, we did observe an asymmetry in the proportion correct for the /a/ signals; overall they were discriminated correctly less than the /u/ and /i/ signals. This asymmetry was due to the signals in the middle of the continuum. This led us to hypothesize that the spectral envelope affected discrimination of the /a/ signals. Clustering, and especially the texture analyses coupled with MDS, resulted in a signal space that was able to group signals mostly along vowel categories. Statistical analysis of the psychophysical responses and analysis of the MDS signal space indicated the most salient attributes of the signals were most likely those previously reported in the literature (spectral envelope, spectral center of gravity, source waveform structure).

The consistency between the psychophysical responses from human observers and the clustering results (using ecologically-motivated signals and signal representations) strongly suggests that basic sensory features and attributes of auditory signals are what primarily determine signal groupings and signal space orientations. The data are not informative as to how cognitive information may contribute, though the majority of clusters and groupings did have relevance to phonetic information (Sinnott et al., 1997). The experimental design could be

expanded on in several ways. First, the psychophysical paradigm could be a complete design, using the schwa vowel as a fourth signal; to make the discrimination efficient, a 2IFC same-different paradigm could be used. Second, a variety of synthetic signals using male, female and child vowel formant values could be employed and compared to investigate the resolution of normalization of speaker differences. Lastly, a greater variety of signals (synthesized in the manner described in this experiment, synthesized using Klatt synthesis, actual vowel tokens) could be transformed using a cochlear representation, analyzed according to texture and then subjected to MDS of the signal similarities and differences

Figures

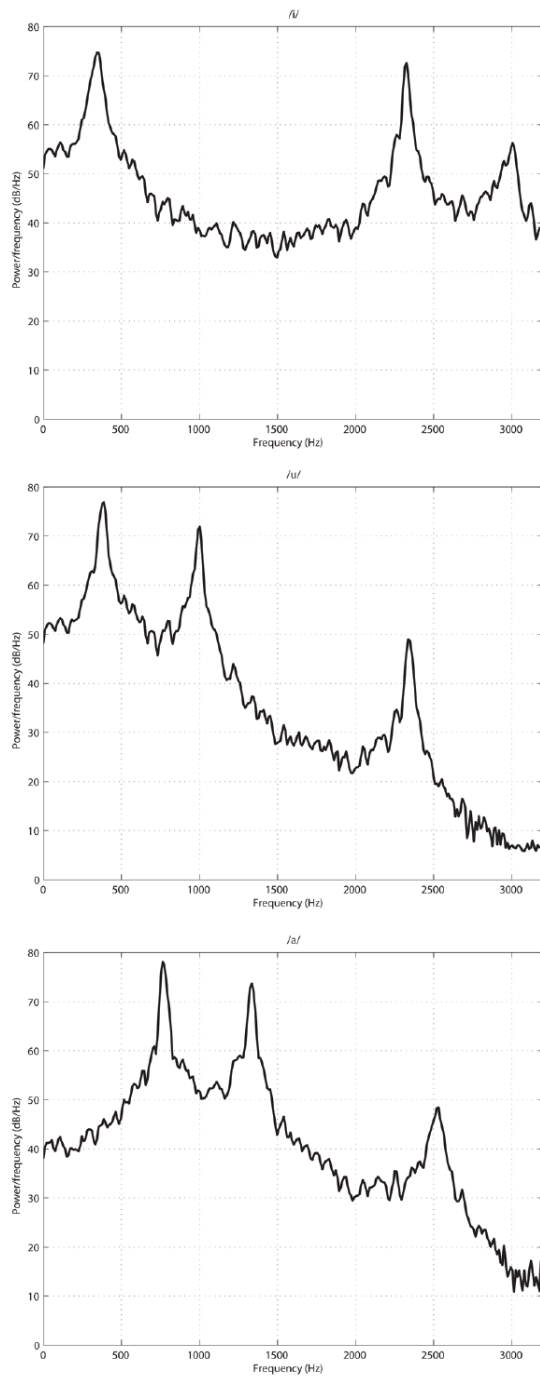


Figure 1. Spectral envelopes of vowel transfer functions employed in the experiment.

Top panel illustrates the transfer function for /i/, middle panel transfer function for /u/ and the bottom panel the transfer function for /a/ (/a/). The x-axis is frequency (Hz);

y-axis is the power spectral density (PSD) estimate (dB/Hz). Power spectral density estimates were computed using Welch's method. The vowels /i/ and /u/ are the high vowels and /ɑ/ and /u/ are the back vowels. The transfer functions are acting as filters on a white noise signal. Transfer functions were created by taking data derived from American English male formant values (Hz) for each vowel (Hillenbrand et al., 1995) and converting the values into the coefficients of a transfer function to filter the various source waveforms employed. The visualization illustrates several important features of each vowel. For the high vowels, it is easy to see that the value of the first formant is relatively close to the fundamental frequency (150 Hz) and that the spectral valley below the first formant is shallow compared to the transfer function for /ɑ/. For the back vowels, the first and second formant vowels are rather close to each other and the spectral tilt is more pronounced than in the front vowel /i/.

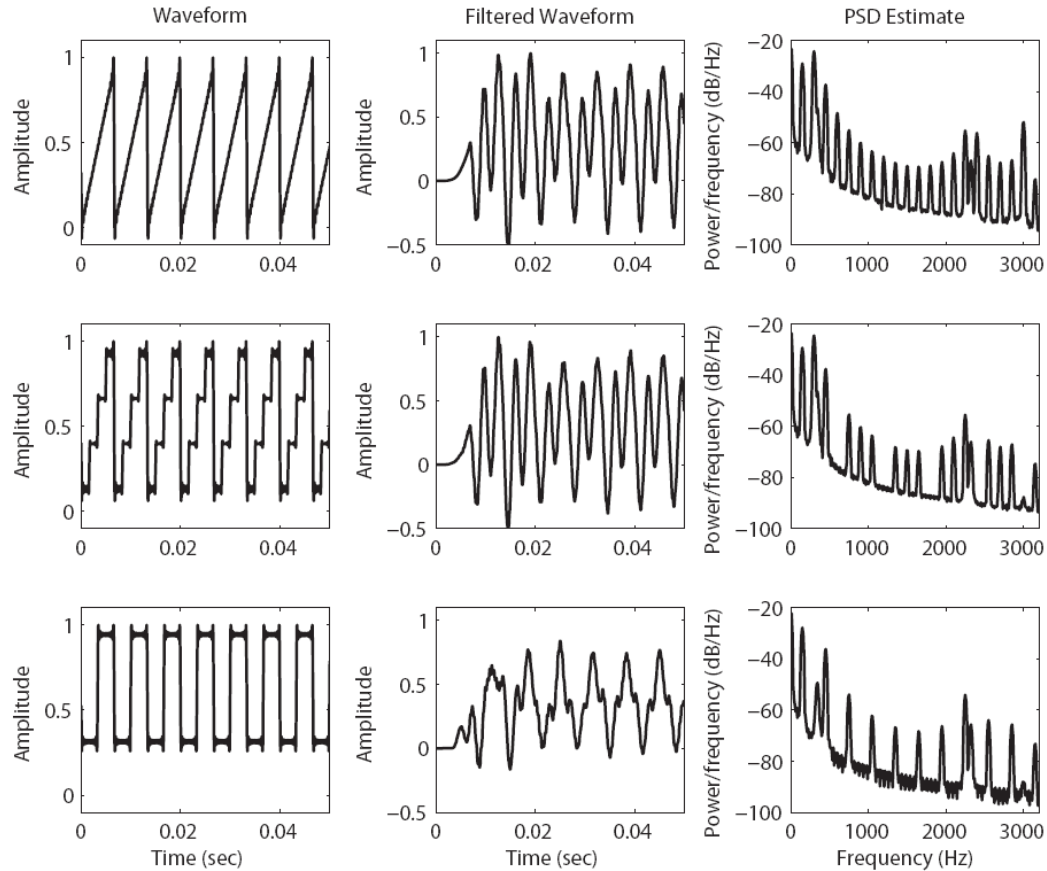


Figure 2. Illustration of stimuli construction, with transfer functions corresponding to the vowel /i/ as an example. Left column illustrates the structures of several source waveforms in the temporal domain; middle column illustrates the structure of the source waveform after filtering by the transfer function; right column illustrates the PSD estimate of the filtered waveform. Top row illustrates the sawtooth wave approximating the glottal excitation pattern; middle row the sawtooth wave with every fourth harmonic removed; bottom row sawtooth wave with every second harmonic removed. Source waveforms were constructed using Fourier synthesis. Visualization of source waveform structure in the temporal domain illustrates (i) the effect of Fourier synthesis in producing the harmonic waveforms and (ii) the effect of selectively removing harmonics to produce discontinuities (in the frequency domain)

and timbre differences between the source waveforms. Visualization of the transfer function (filters) on the source waveforms illustrates the steady-state temporal dynamics of the signals. The steady-state nature of the signals is readily apparent, i.e., the waveforms oscillate at a combination of the fundamental frequency and the harmonics present. Though somewhat difficult to see, close inspection of the sawtooth wave approximating the glottal excitation pattern and the sawtooth wave with every fourth harmonic removed reveals that (i) the essential sawtooth nature of the signals is preserved, despite differences in the harmonic structure and (ii) the selective removal of harmonics results in slight differences in the temporal evolution of the signals. PSD estimates more clearly reveal the timbral differences in frequency domain. The spectral peaks that stand out most prominently are at the values of the formants; though synthetic the signals employed exhibit the comb structure typical of vowel tokens and vowel-like signals. The effect of selectively removing harmonics is also clear; there are ‘gaps’-- discontinuities introduced by the removal of harmonics -- in the spectrum where those harmonics are not present.

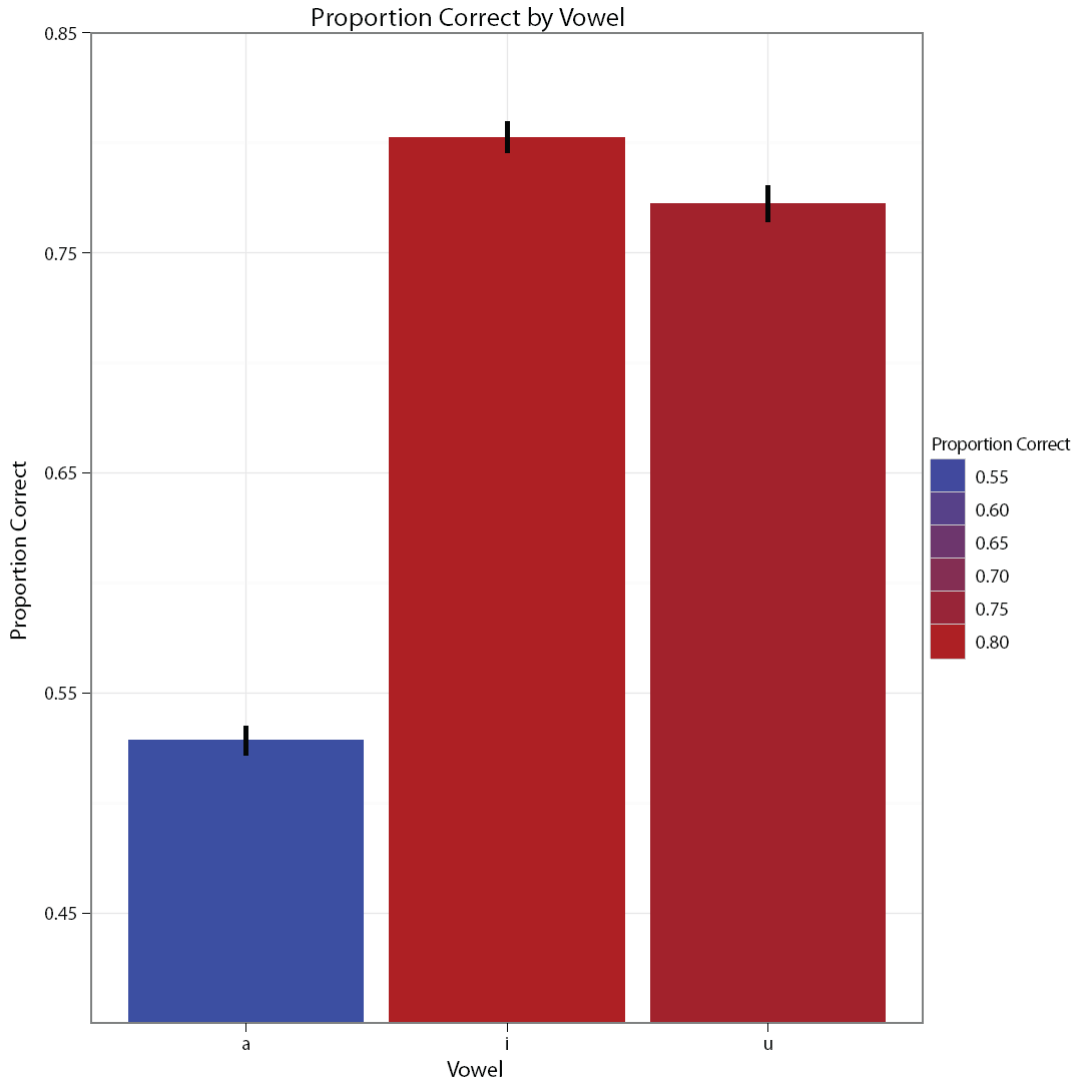


Figure 3. Proportion correct for each vowel signal class, pooled experimental data. X-axis plots each vowel, y-axis total proportion correct over all trials. Data are taken from pooled data after rejecting participants that performed at chance or worse (< 50% correct) over all items. The legend indicates the proportion correct via a color scale; colors towards the blue end of the spectrum indicate greater incorrect responses. There is a clear asymmetry in the proportion correct between the /i/ and /u/ signals and the /a/ (/a/) signals. Over all trials, the /i/ and /u/ signals were

discriminated more correctly (~0.80) than the /a/ signals (~0.50). The asymmetry observed may be a result of the formant structure of the /a/ signals.

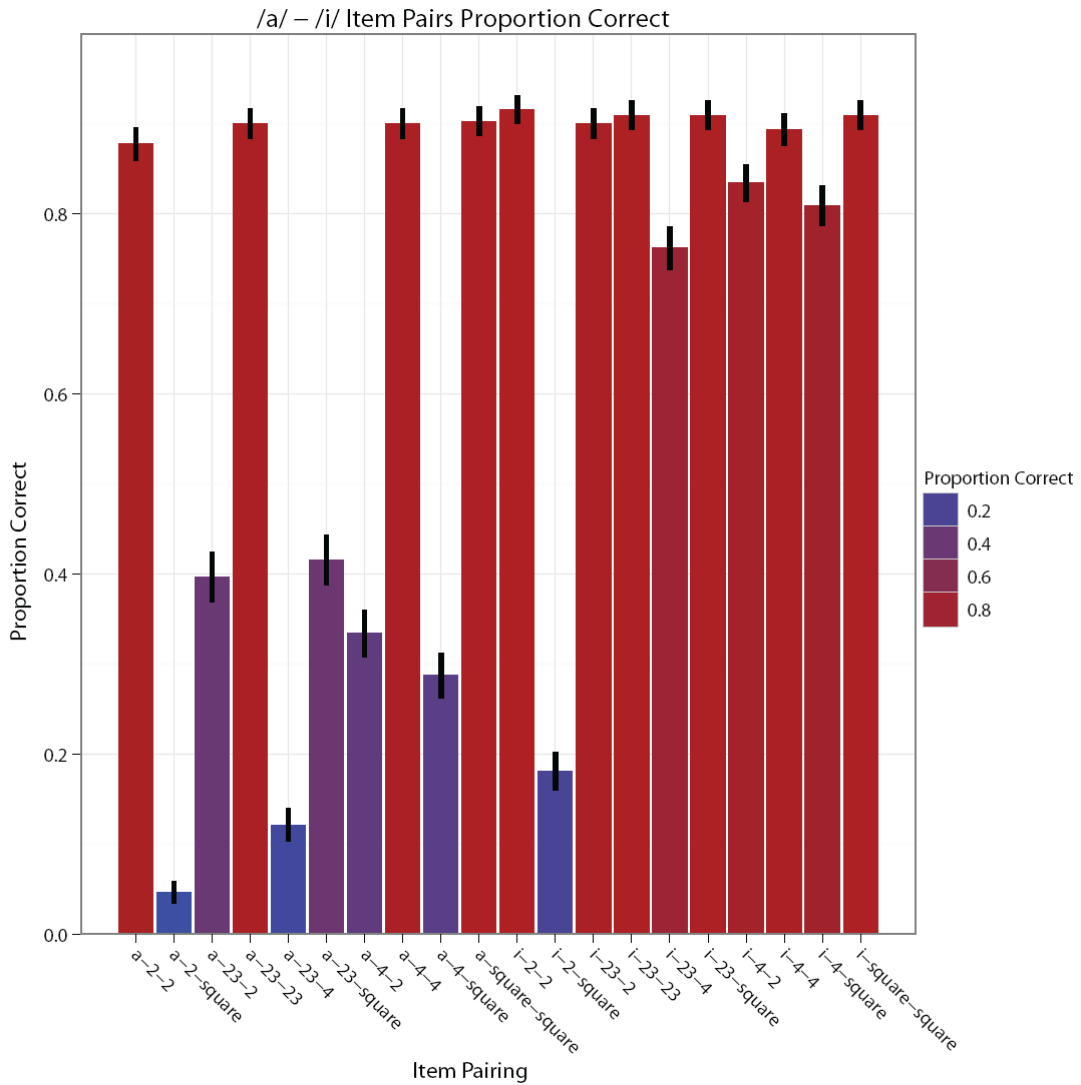


Figure 4. Proportion correct for each item (trial) across participants for the /a/ - /i/ experimental assignment. Y-axis is proportion correct, x-axis indicates the trial. X-axis labels are formatted as follows: the letter at the beginning of the label indicates the vowel signal, numbers indicate harmonic structure of source waveform (i.e. '23' =

sawtooth wave comprised of 23 harmonics, '4' = sawtooth wave with every fourth harmonic removed, '2' = sawtooth wave with every second harmonic removed, 'square' indicates square source waveform). The legend colors indicate overall proportion correct as in Figure 3. In line with our hypotheses, the trials containing the metameric comparison (a-2-square and i-2-square) were incorrectly discriminated the majority of the time. There is also an obvious asymmetry between the vowel signals. The /i/ signal trials where the metamers were not compared exhibited high proportion correct values, with the minimum being ~ 0.76 . In contrast the /a/ signal trials that were not either the metameric comparison or where the signals being compared were identical, exhibited low proportion correct values. Values for these trials ranged from ~ 0.15 to 0.4.

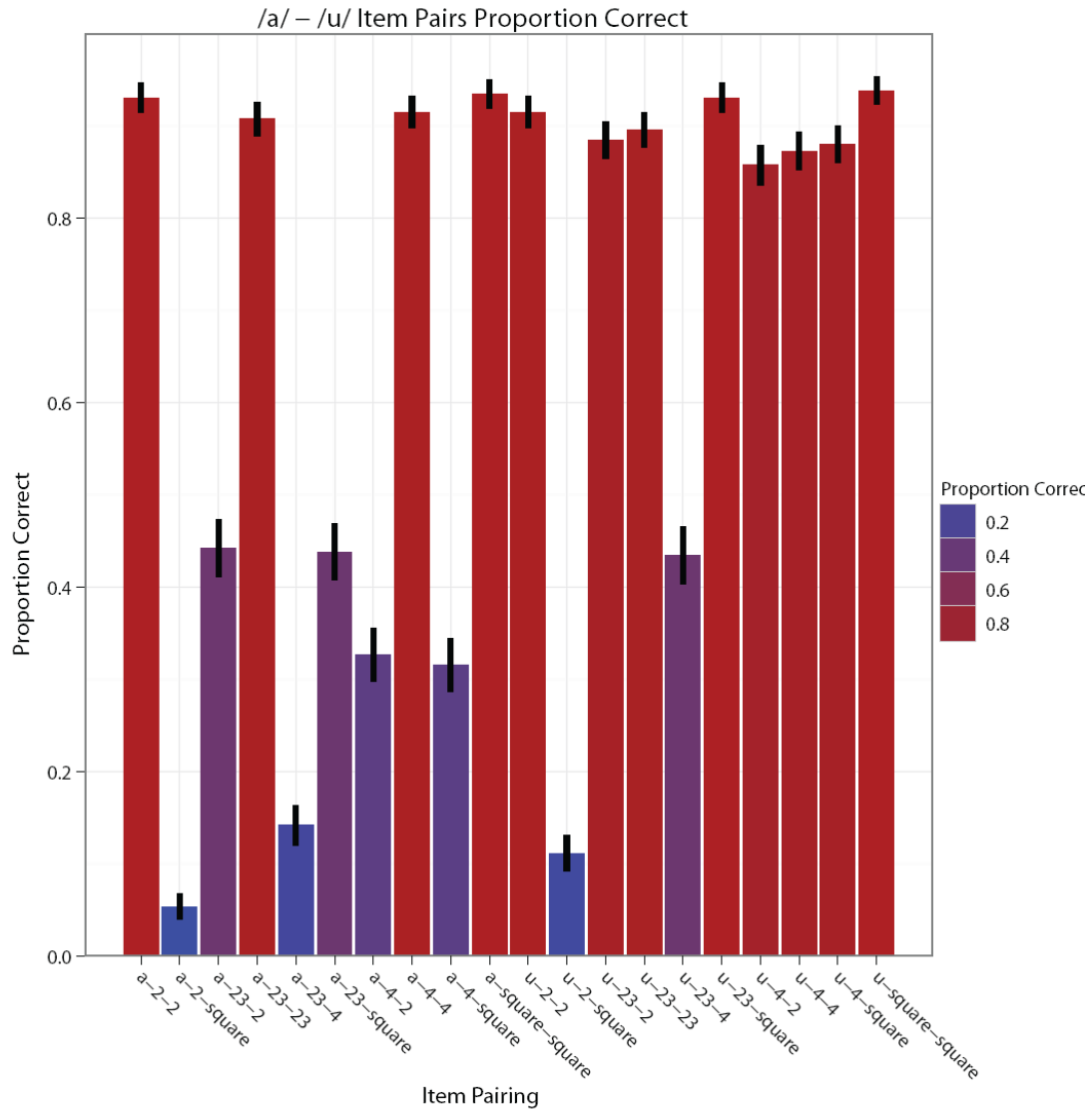


Figure 5. Proportion correct for each item (trial) across participants for the /a/ - /u/ experimental assignment. Conventions are the same as used in Figure 4. As in the /a/ - /i/ assignment, our hypotheses concerning metameric comparison were supported by the data; the trials containing the metamers exhibited very low proportion correct values. The asymmetry between the /a/ and the /u/ signals was also present as in the /a/ - /i/ assignment, with the /a/ signal trials that were not either the metameric comparison or where the signals being compared were identical having low

proportion correct values, ranging from ~ 0.15 to 0.45. The results of the proportion correct analyses in both pools suggests that differences in formant structure between vowel transfer functions gives rise to these asymmetries (see *Discussion*). The signals in the middle of the continuum were thus responsible for the overall proportion correct asymmetry observed.

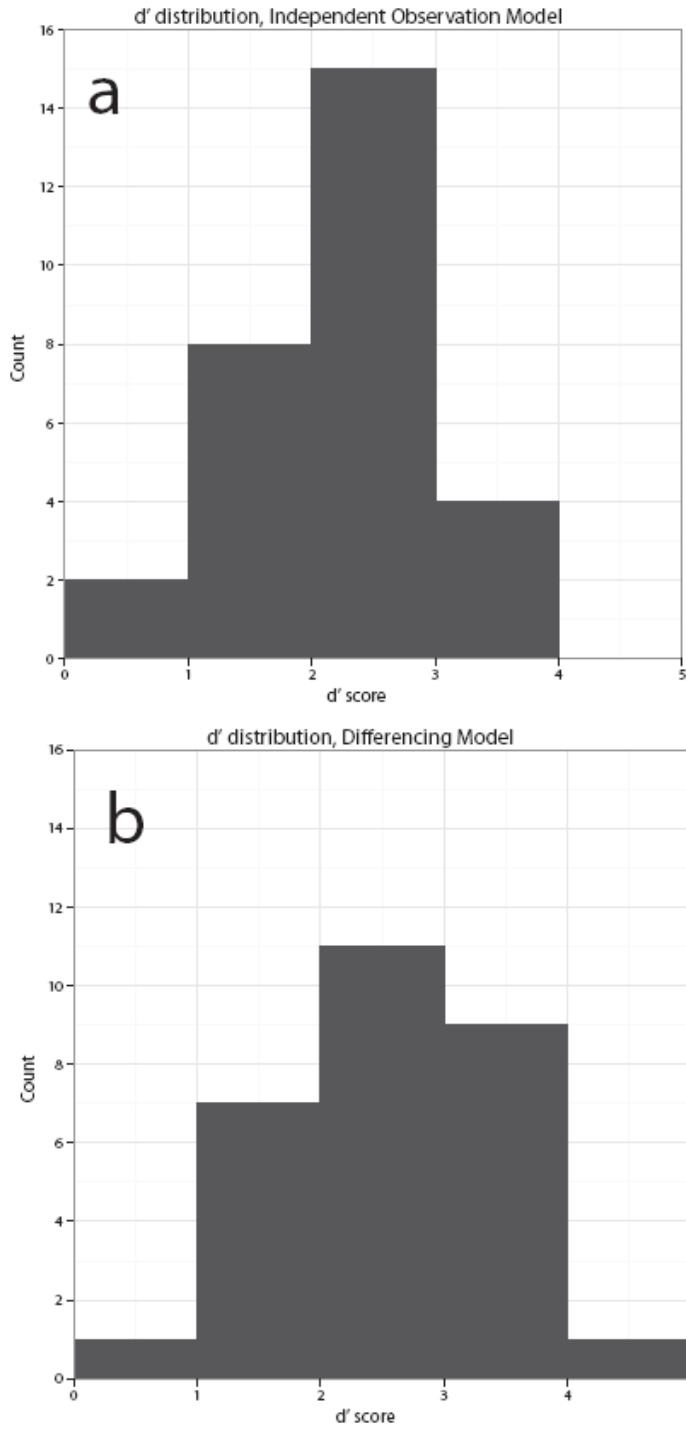


Figure 6. d' (observer acuity) score distributions for the independent observation (a) and differencing models (b) for the 1AFC same-different experimental paradigm. Data are taken from the pooled experimental data, after rejection of participants who

performed at chance or worse over all trials. The distribution of observer acuity distributions for each model demonstrates that overall, participants were able to discriminate between signals rather well. For the independent observation model, the majority of d' scores were in the range from 1 to 3, while for the differencing model, the majority of scores were in the range from 1 to 4.

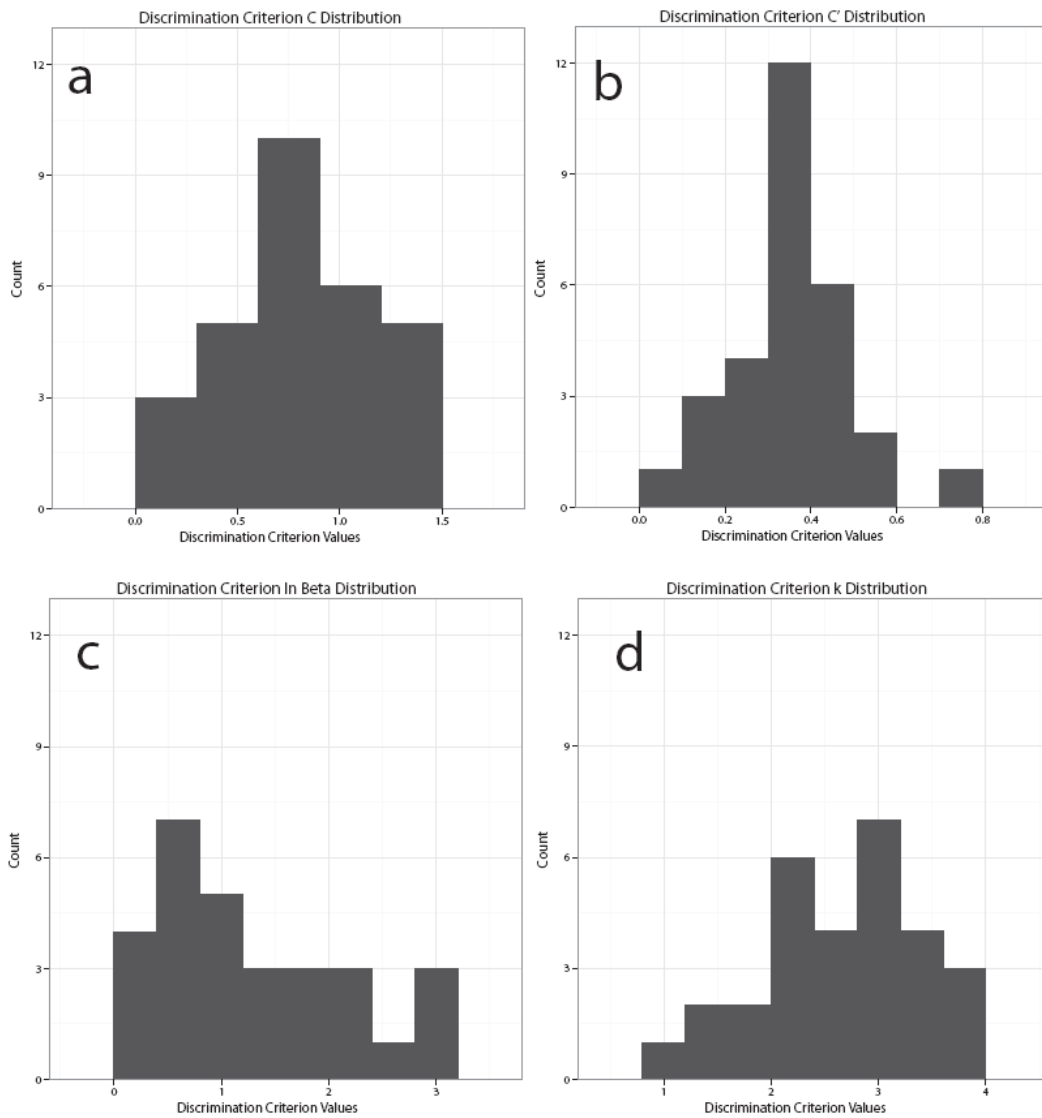


Figure 7. Figure 7 illustrates the distributions of the different discrimination criteria for the independent observation (C , C' , $\ln \beta$) and differencing models (k). Figure 7a illustrates the distribution for discrimination criterion C , 7b for criterion C' , 7c for $\ln \beta$ and 7d for k . The distributions for the different criteria vary greatly, possibly indicating a wide variety of perceptual dimensions the various participants used in discriminating between signals. The only two discrimination criteria to exhibit a strong correlation to d' scores were $\ln \beta$ (0.808) and k (0.794).

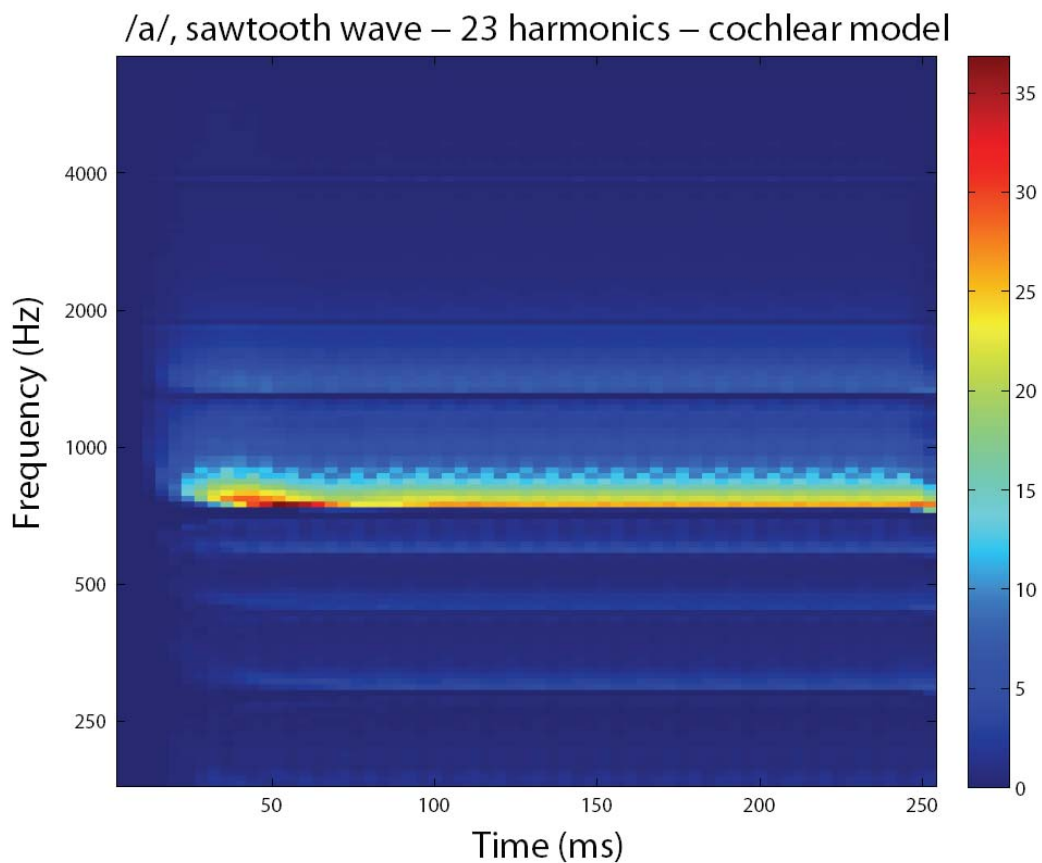


Figure 8. Figure 8 illustrates the cochlear model representation of an experimental signal. The signal is the sawtooth wave approximating the glottal excitation pattern

for the vowel /a/. The y-axis is frequency (Hz) and the x-axis is time (msec). Color values indicate relative intensity; red values indicate a high concentration of energy. Several aspects of the signal are apparent from the visualization of the cochlear representation. First, since the signal is steady-state, then there is little temporal evolution; what temporal evolution is present comes from the ramping of the signals at onset and offset and from the initial effects of the transfer function on the source waveform. Second, the comb structure typical of vowels and vowel-like signals is present; several horizontal bands are present, indicating the presence of the harmonics. These bands appear rather faint due to spectral tilt. Third, a large amount of energy is present in the region of the first two formants (~700-1000 Hz); the peaks in the spectral envelope arising from the first two formants are not present, but rather they are averaged in a sense to produce a frequency range where a substantial amount of energy is concentrated.

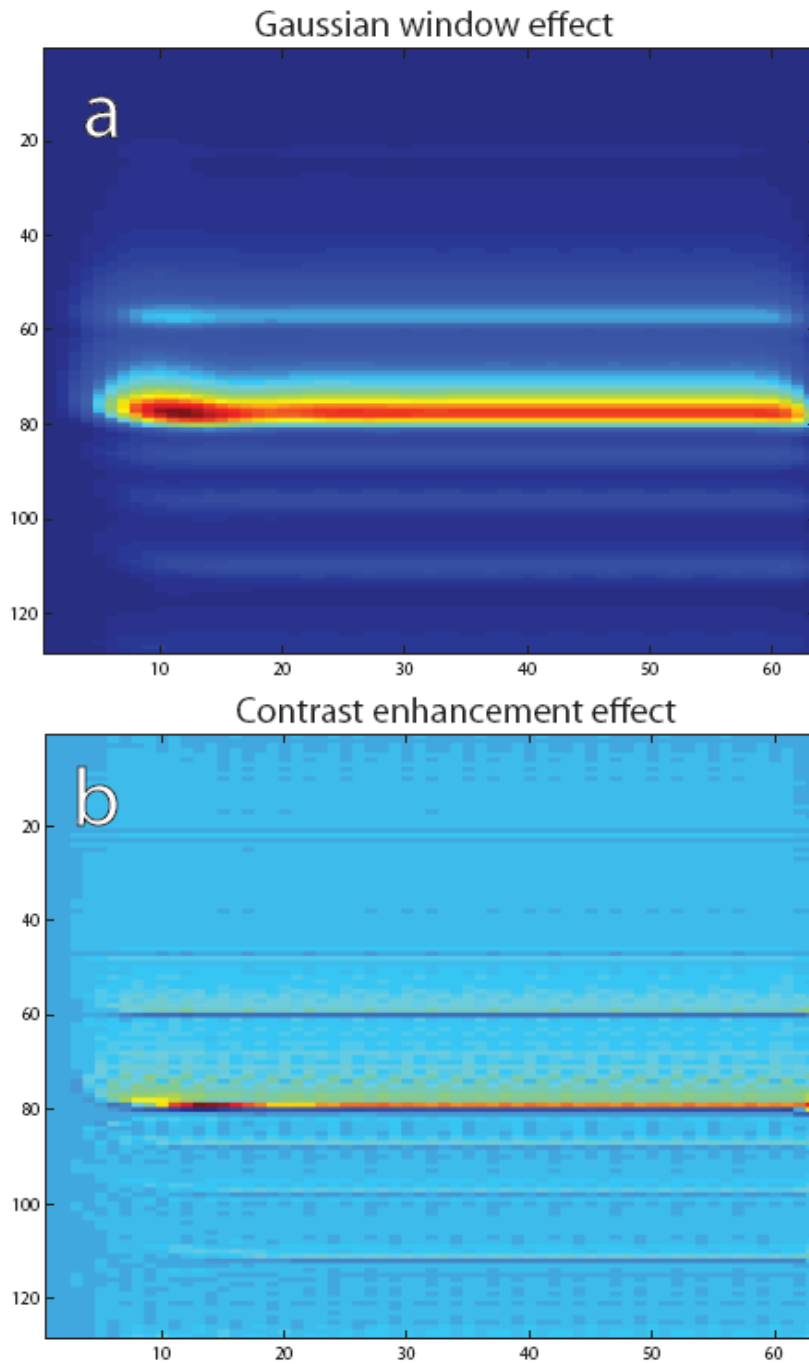


Figure 9. Figure 9 illustrates the effects of the different filters on the cochlear representation of the signal. The signal and conventions used for the colormap are the same as in Figure 8; the axes are the indices of the matrices of the cochlear representation. The Gaussian filter (9a) can clearly be seen to emphasize the gross

features of the signal, with several crucial differences from the visualization provided in Figure 8. First, the amount of energy concentrated in the vicinity of the first two formants is more clearly pronounced and is more evident throughout the duration of the signal. Second, the presence of the third harmonic is more clearly visible. Third, the contribution of the source waveform harmonics is attenuated by employing the Gaussian smoothing filter. Figure 9b illustrates the effect of the contrast enhancement (negative of Laplacian) filter on the cochlear representation. While the energy concentration in the range of the first two formants is still visible, contrast enhancement more clearly emphasizes the harmonic structure of the source waveform. In fact, the contribution of harmonics greater than the third formant in spite of spectral tilt is more readily apparent.

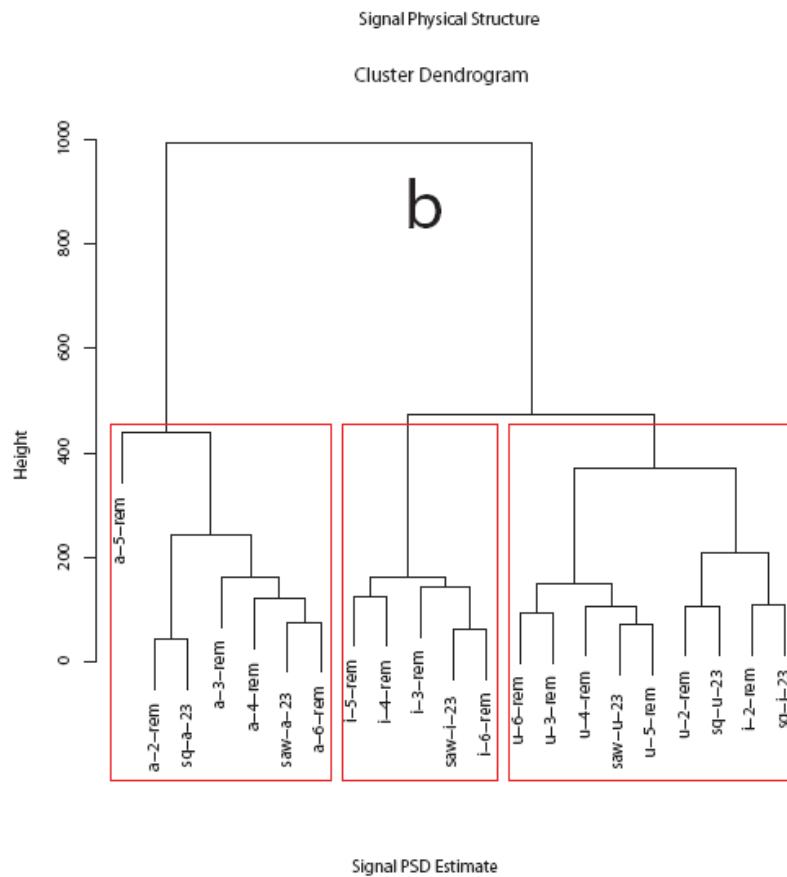
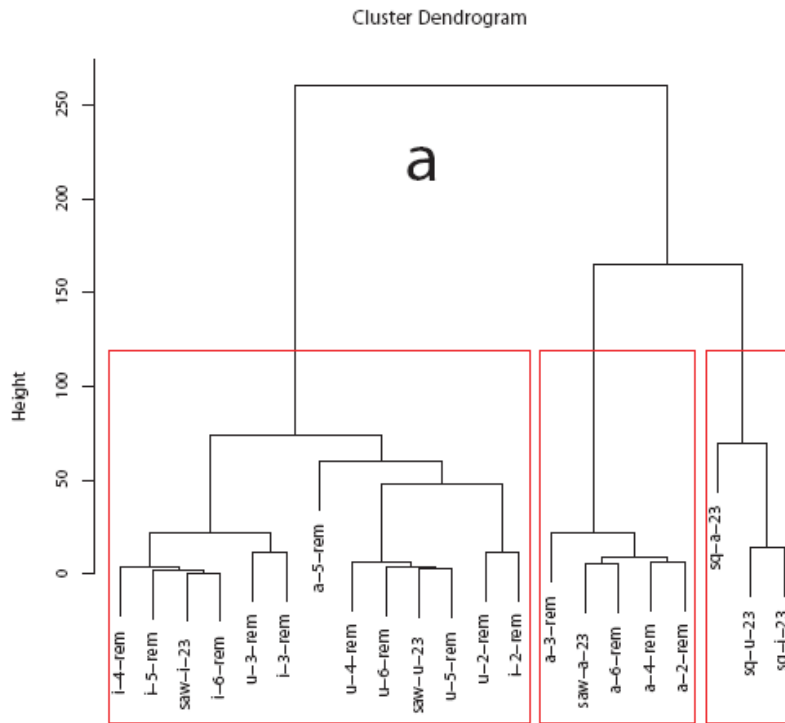
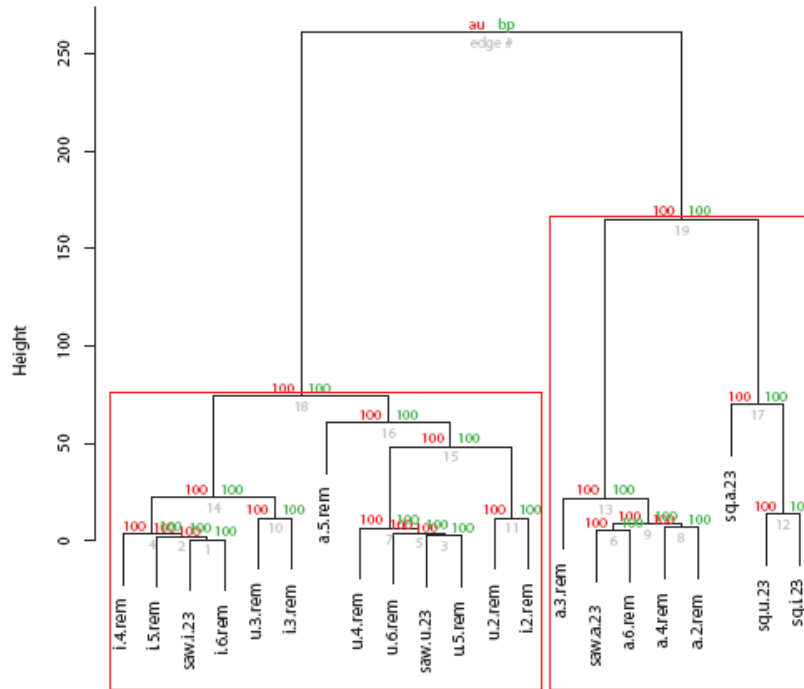


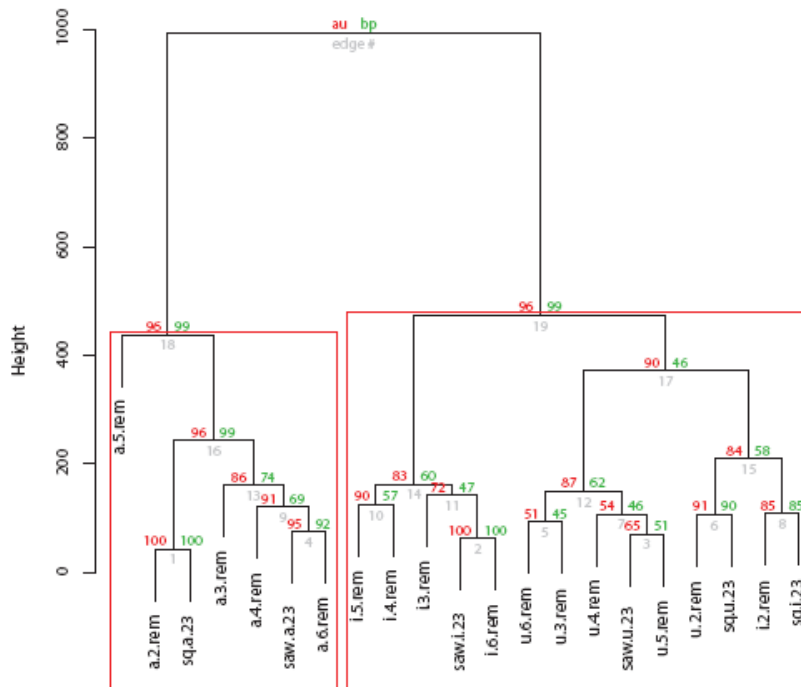
Figure 10. Dendrograms for physical structure (waveform) and PSD estimates, initial 3 cluster solution using Ward's method and Euclidean distance. Figure 10a displays the dendrogram for the waveform physical structure. The three clusters largely group the signals by vowel class as well as source waveform; the high vowel signals (/i/ and /u/) are in a separate cluster from most of the /a/ signals, the square source waveform signals are located in a separate cluster. Figure 10b displays the dendrogram for the PSD estimates. For this dendrogram and data representation, the signals are grouped mostly according to vowel category, with signals from each vowel mostly being placed within separate clusters.

Cluster dendrogram with AU/BP values (%)



Signal Physical Structure

Cluster dendrogram with AU/BP values (%)



Signal PSD Estimate

Figure 11. Bootstrap supported dendrogram clusters, 1000 bootstrap resamplings. The numbers in red indicate approximately unbiased p-values; numbers in green are the bootstrap probability values. The number of bootstrap supported clusters was less than both the clusters from our initial hypotheses and those indicated by the within sum of squares. Figure 11a displays the dendrogram for the physical structure; signals were mostly clustered according to vowel category (high or low); the square source waveform signals were clustered along with the /a/ signals. Figure 11b displays the dendrogram for the PSD estimates; the signals are placed in two separate clusters with the high vowel signals in a separate cluster from the low vowel signals.

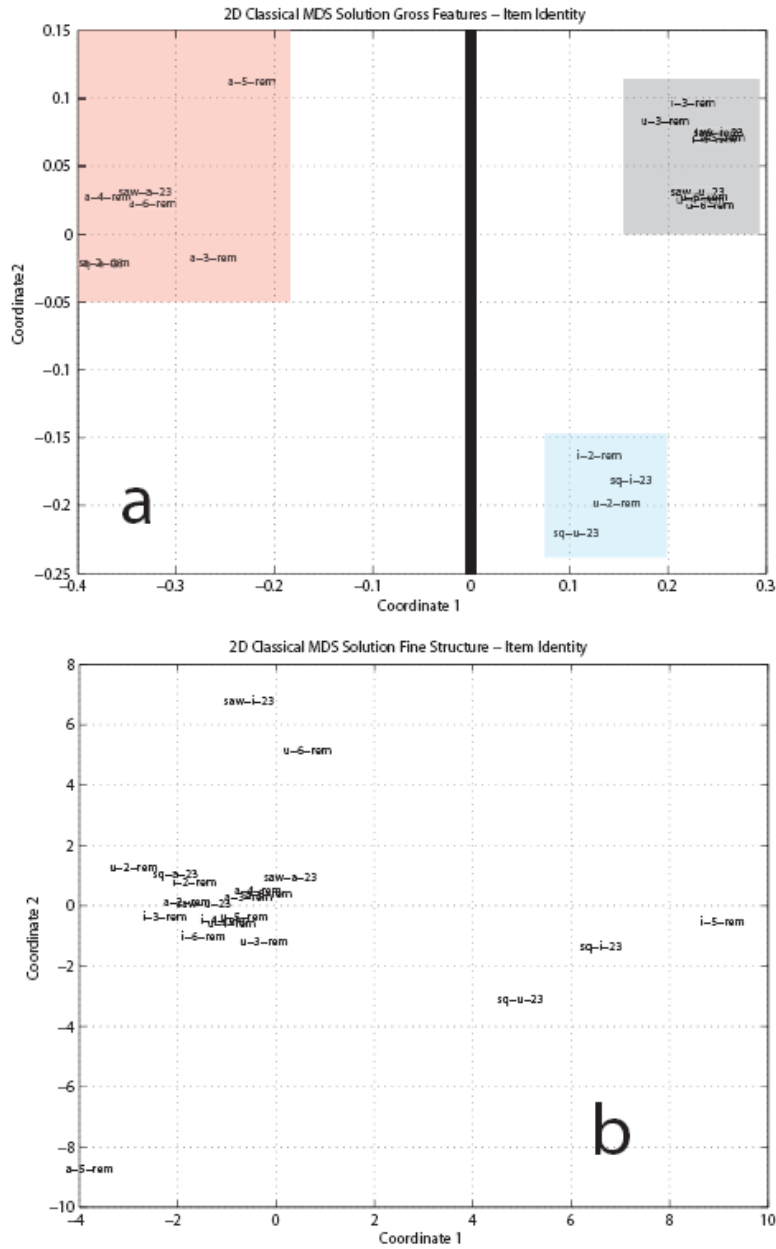


Figure 12. Two-dimensional MDS solutions for the cochlear representations. The first coordinate is on the x-axis, second coordinate on the y-axis. Text strings indicate the placement of the signals within the MDS space. Figure 12a displays the 2D MDS solutions for the gross feature comparison; 12b for the fine structure comparison.

Coordinate 1 seems to index vowel height; the high vowels are located on the right side (values > 0 in Coordinate 1, to left of black line) of the MDS space. The second Coordinate does not seem to index or group vowel information, but does seem to group the high vowel metamers from the rest of the signals. For the gross feature comparison, there are three distinct areas within the MDS space, delineated by rectangles (left of solid black line). One area contains the metamers for the high vowel signals (blue rectangle), located in the range (0.05, 0.2) for Coordinate 1 and (-0.23,-0.15) for Coordinate 2. A second area contains the /a/ signals (red rectangle); within this area, the metamers are mapped to the same space, approximately at the coordinates (-0.4,-0.03). A third area contains the non-metamer high vowel signals (gray rectangle). Unlike the gross feature comparison MDS space, the fine structure comparison across vowel categories exhibits no ordered structure within the space.

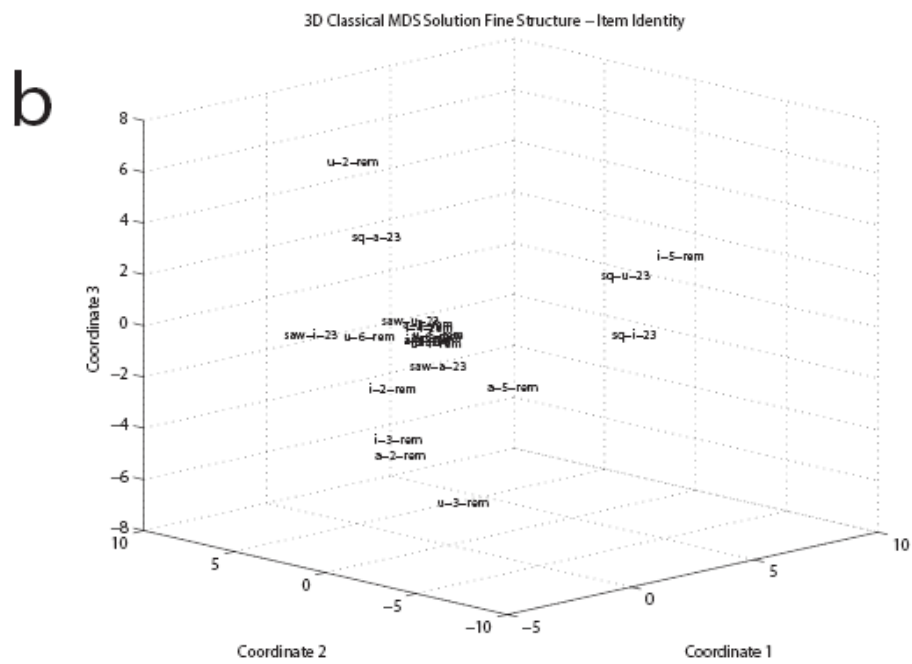
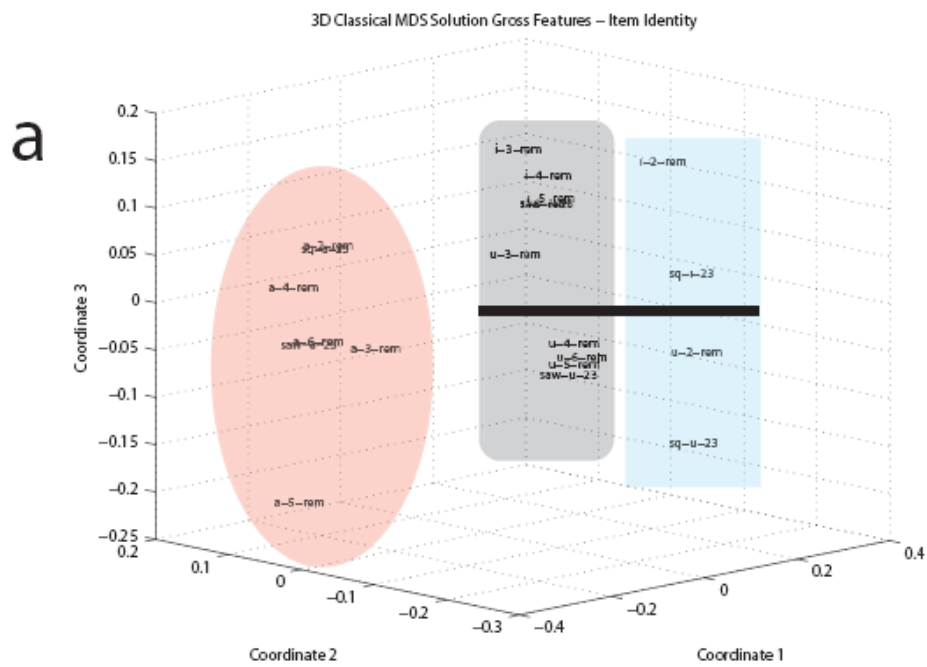


Figure 13. Three-dimensional MDS solutions for the cochlear representations. Figure 13a visualizes the 3D solution for the gross feature comparisons; 13b for the

fine structure comparisons. The first two coordinates are unchanged from the 2D solution. The 3D plot viewed from this angle gives a sense of the density of the MDS space for this dimensionally reduced solution. The /a/ portion of the space (red ellipse) is considerably less dense than the spaces where the /i/ (gray rounded rectangle above black line) and /u/ signals (blue rectangle below black line) are clustered; the /a/ with a source waveform where every fifth harmonic is removed is still considerably further away than the rest of the signals. The /a/ metamers and the /a/ with every sixth harmonic removed and approximating the glottal excitation pattern are closely positioned in the three dimensional solution. High vowel non-metameric signals are highlighted by the gray rectangle; high vowel metamers are highlighted by the blue rectangle. The /i/ signals are positioned ‘higher’ (further up in elevation on the plot, division indicated by solid black line) than the /u/ signals; this may be analogous to tongue position for these signals. The relationship holds for the /i/ and /u/ signals overall, including the metamers, with a single exception (u-3-rem). The three-dimensional solution for the fine structure comparisons did not yield a high GOF nor was it apparent what each coordinate indexed.

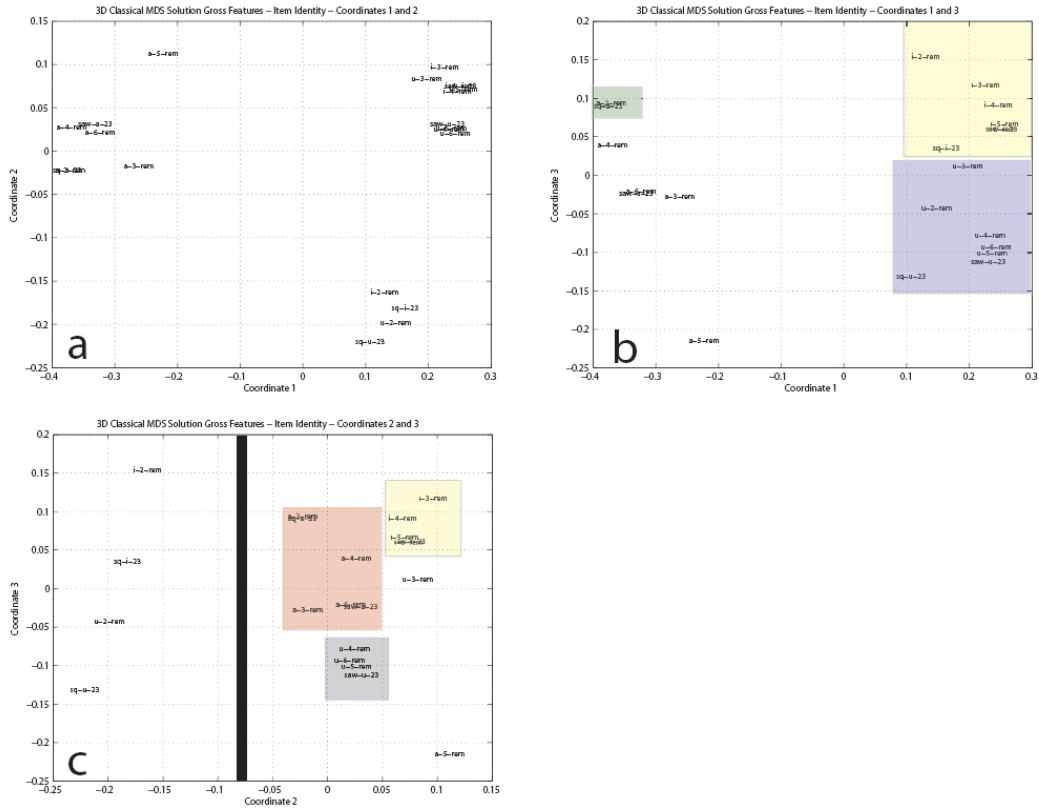


Figure 14. Three-dimensional MDS solution visualization for the gross feature comparison, with each coordinate plotted against each other in separate planes. Figure 14a visualizes the space for coordinates 1 and 2; 14b for coordinates 1 and 3; 14c for coordinates 2 and 3. The visualization of coordinates 1 and 2 in Figure 14a is unchanged from Figure 12. In the visualization of the plane where coordinates 1 and 3 intersect, it is apparent that coordinate 3 separates the high vowel signals according to their vowel categories, i.e. /i/ (yellow rectangle) and /u/ (blue rectangle). There is no apparent systematic ordering of the /a/ signals along the third coordinate axis. However, the /a/ metamers (green rectangle) are close together in the MDS space, in line with our hypotheses. Visualization of the 2-3 plane demonstrates the second coordinate separates the high vowel metamers from the rest of the signals (division

denoted by black line). This plane appears to index a combination of vowel category and spectral information; the majority of /a/ (red rectangle), /i/ (yellow rectangle) and /u/ (blue rectangle) signals are clustered together and progress along coordinate 3 seems to index spectral complexity or harmonic structure. Only two signals (aside from the metamers) were not included in their vowel category in this plane.

Preattentive Classification and Physiological Measurement of Ecologically Approximate Synthetic Signals using MEG

Introduction

It is a remarkable fact about the human auditory and communication systems that a healthy listener -- one without hearing loss, stroke, aphasia, autism, etc. -- is able to successfully group auditory signals into relatively abstract categories within fractions of a second. For example, it has been argued on the basis of event-related potential (ERP) data that within 50-100 msec after stimulus onset, listeners can preconsciously distinguish between living and inanimate objects in the auditory domain (Murray et al., 2006). Apparently then, for successful auditory cognition to occur, and thus successful communication, the brain exploits subtle cues and differences in signals to group them into categories. However, we understand surprisingly little about how short auditory segments elicit the complex internal representations associated with auditory cognition in general and speech in particular. Despite there being a considerable body of work regarding cochlear function, signal transduction through the auditory pathway, cortical parsing of auditory signals and linguistic effects on sound perception, we still lack a complete understanding of how vastly disparate auditory patterns are learned, stored, retrieved and classified.

Though it is unclear how the brain creates these functional groupings of auditory signals, evidence from psychophysical and neuroimaging studies have identified several dimensions that are most salient for determining how a signal is grouped/evaluated. In addition to an increased understanding of the physical cues used in creating functional groupings, there is a solidifying consensus concerning the existence of auditory processing pathways, similar to those in the visual system, that process the spatial location ('where') and identity ('what') of the signal encountered (Belin and Zatorre, 2000; Inui et al., 2006; Okamoto et al., 2010). Imaging studies using electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) have identified cortical areas that specialize in the recognition of auditory objects as well as selective discrimination of speech and voice recognition (Belin et al., 2000; Hickok and Poeppel, 2000; Belin and Zatorre, 2003; Belin et al., 2004; Obleser et al., 2006; Hickok and Poeppel, 2007). The auditory system is also structured in such a way as to take advantage of the objects encountered in the environment and ecological communication signals. For example, computational models derived from physiological measurements can capture the harmonic structure found in vocalizations (Shamma and Klein, 2000; Lewicki, 2002; Smith and Lewicki, 2006) as well as the broadband carrier statistics of vocal carrier signals (Moore, 2003; Diehl, 2008; Mesgarani et al., 2008) and the time courses of these events over a wide range of situations and signal presentations (Loizou et al., 2003). The auditory percept most responsible for identification of a signal as it is processed by the auditory pathway is timbre, or tone color (Fellows et al., 1997; Belin et al., 2000; Binder et al., 2004; Zatorre et al., 2004). In regards to the ecological speech signal,

the excitation of the vocal folds and speaker vocal tract shape determines the timbre of segments produced (generally phonemes). To increase understanding of how functional groups are created then, it would be best to study timbre and its relationship to other basic auditory attributes such as pitch and loudness.

Despite timbre being a “psychologically relevant” auditory cue involved in identifying the source of a signal, its physiological and psychological bases are not well understood. This is due mostly to the multidimensionality of timbre. Timbral cues arise primarily from spectral (McAdams et al., 1995; Caclin et al., 2005; Caclin et al., 2006), envelope amplitude modulation (AM), frequency modulation (FM) (Hall, 1977; Iverson and Krumhansl, 1993; Meyer et al., 2006; Luo et al., 2007) and contextual information (Bigand and Pineau, 1997; Pardo and Fowler, 1997; Nelken, 2004; Obleser et al., 2007; Obleser and Kotz, 2010); these cues often interact with one another at very fine time scales, making investigation quite complicated (Melara and Marks, 1990; Diesch and Luce, 1997; Caclin et al., 2007; Caclin et al., 2008). However, use of neuroimaging techniques can help in discovering the relationship between timbre perception and auditory architecture (Crummer et al., 1994; Meyer et al., 2006; Caclin et al., 2008). Current neuroimaging evidence links the results of the dimensions commonly found in psychophysical studies -- attack time (Singh and Hirsh, 1992; Samson et al., 1997; Singh and Bregman, 1997; Caclin et al., 2005; Caclin et al., 2007), spectral centroid/center of gravity (Grey and Gordon, 1978; Iverson and Krumhansl, 1993; Lakatos et al., 1997), spectral flux (Iverson and Krumhansl, 1993; McAdams et al., 1995; Caclin et al., 2005; Caclin et al., 2008),

spectral spread (Marozeau et al., 2003) and ‘brightness’ (von Bismarck, 1974b, a; Grey and Gordon, 1978; Kendall and Carterette, 1993; Pressnitzer and McAdams, 1999; Marozeau and de Cheveigné, 2007) -- to quantifiable physiological measures such as the latency and amplitude of evoked response components and brain areas sensitive to specific features of an auditory signal/object. Basic physical features of auditory signals also contribute to timbre perception, as different frequencies and waveform shapes tend to have specific semantic descriptions of their timbre (von Bismarck, 1974b, a; Hartmann, 1998). The dimensions commonly reported in the literature tend to be the most salient perceptually (depending on signal context and task demands) and generally have a basis in the physical structure of the signal (e.g. spectral envelope, fundamental frequency). Further confounding the investigation of timbre perception is the fact that, to some extent, timbre is not separable from other attributes such as pitch and loudness (Melara and Marks, 1990). For example, sine waves and piano tines are described as having a ‘pure’ sound, while fundamental frequency affects if a sound is described as being ‘dull’ or ‘buzzy’ (Hartmann, 1998; Marozeau and de Cheveigné, 2007).

In relation to the ecological speech signal, timbre contributes to processes such as speaker identification, gender and affective state (Fellows et al., 1997; Belin et al., 2000; Binder et al., 2004; Zatorre et al., 2004). This is most likely accomplished by the auditory system's evaluation of the statistics of a signal: its underlying dynamics (Balzano, 1986; Lakatos, 2000), commonalities in signal structure and fate arising from the sound sources processed by the cochlea (Bregman

and Pinker, 1978) and inferences about sound source (Lakatos et al., 1997). Studies of musical timbre have demonstrated that a signal's timbre is not determined only by its own spectral profile, but by also that of the signals preceding and subsequent to it (Bregman, 1990; Deutsch, 1999). Processing of ecological signals in humans encompasses not only general auditory processing areas, but also areas that seem to process different aspects of speech and music preferentially, forming a distributed analysis network (Samson and Zatorre, 1994; Auzou et al., 1995; Hickok and Poeppel, 2000; Samson et al., 2002; Samson, 2003; Hickok and Poeppel, 2007; Obleser et al., 2008). Evidence for the hemispheric specialization for different aspects of general auditory processing and speech and music perception has been provided via neuroimaging and lesion studies using preattentive/preconscious signal evaluation as well as psychophysical measures (Boucher and Bryden, 1997; Chait et al., 2004; Jacobsen et al., 2004b; Jacobsen et al., 2004a).

Electroencephalographic and magnetoencephalographic (MEG) studies have demonstrated that perceptual and physical attributes of auditory stimuli are reflected in the timing and field deflections of the evoked components. The most prominent components are the P50/M50, N100/N1m/M100 and P2/P2m, with each component showing differential sensitivity to various aspects of an auditory signal. These aspects include signal bandwidth, fundamental frequency, waveform shape (Roberts et al., 2000) and signal level (Jenkins et al., 2010). For example, the latency and amplitude of the M50 is sensitive to signal bandwidth (Chait et al., 2004). The N1m/M1000 is well-studied and has been shown to be sensitive to perceptually

primary and salient parameters such as pitch, frequency, vowel information, temporal modulation of auditory signals, spectral complexity and is additionally modulated by musical and frequency discrimination training (Eulitz et al., 1995; Diesch et al., 1996; Poeppel et al., 1997; Diesch and Luce, 2000; Gage and Roberts, 2000; Roberts et al., 2000; Brattico et al., 2003). The P2/P2m evoked component seems to index spectral complexity, with spectrally complex signals eliciting greater amplitudes (Shahin et al., 2005; Shahin et al., 2007).

With regards to ecological signals, vowel spectra (Tiitinen et al., 2004, 2005), tokens recorded from human speakers (Tiitinen et al., 2005) and vowel-like sounds (Vihla et al., 2000; Vihla and Salmelin, 2003; Jacobsen et al., 2004a) have been employed as stimuli in a number of studies, with their physical and perceptual characteristics being reflected in different evoked components. Signals of this nature are valuable as experimental stimuli as they have a well-defined spectral structure, have a relatively steady-state nature and tend to elicit robust neuronal activity and are quite salient perceptually. The evoked responses measured using vowel stimuli are modulated by factors such as the structure of the source signal, the perceived pitch of the vowel and the formant structure (spectral envelope) of the stimuli. Data from MEG and EEG experiments also indicate that linguistic information is integrated into the generation of an evoked response (Eulitz et al., 1995; Diesch et al., 1996; Diesch and Luce, 1997, 2000; Obleser et al., 2003b; Obleser et al., 2003a, 2004; Tavabi et al., 2007) and that classification of a sound into its functional group can also be accomplished very rapidly (Murray et al., 2006). More specifically, different classes

of vowels contain certain characteristics that allow them to be classified and distinguished from one another. For example, high vowels (e.g., /i/ and /u/, called 'high' due to position of the tongue in the mouth) tend to have a low-frequency first formant (F1) relative to the low vowels, that is close to the fundamental frequency of the speaker and have damped oscillation that at the frequency of the first formant. Low vowels (e.g., /a/) exhibit large spacing between the first formant and the fundamental frequency and have a deeper spectral valley below the first formant. This results in a comb structure for signals of this nature, and the resonances (formants/spectral envelope) are the primary cue to vowel category (see Figure 1). These properties can be described as phonation and articulation (Tiitinen et al., 2005) and are related to the source and filter characteristics of vocal fold excitation and formant structure, respectively (Fant, 1980; Fant et al., 2000). Vowels and vowel-like sounds, due to their salience, spectral structure and ability to be parametrically controlled aid in the investigation of timbre due to their (i) source waveform structure, (ii) bandwidth effects, (iii) differences in spectral envelope and (iv) temporal modulation of auditory signals.

To investigate how sounds are classified (categorized) preconsciously (preattentively), we employ sounds of intermediate complexity that approximate ecological signals and incorporate some of the dimensions previously reported to be especially salient in the timbre literature. More specifically, we employ parametrically controlled synthetic signals that allow us to probe an ecologically relevant signal, but also timbre perception in a more general sense. We model the

production of acoustic speech in accordance with source-filter theory, where a source signal produced by an object is filtered by a transfer function, giving rise to its specific timbre. As filters, we use formant values (and the transfer functions created by them) of the American English vowels /a/, /i/ and /u/ (with formant values obtained from Hillenbrand et al., 1995). To increase the ecological validity of the signals, the source waveform is a sawtooth wave comprised of twenty-three harmonics (Cleveland, 1977). Creation of a timbral continuum, is accomplished by selectively removing harmonics (from every sixth to every second); the last source waveform is a square wave where the maximum harmonic is the twenty-third. The sounds employed in this experiment differ from real vowel tokens in that they are completely steady-state and as such, lack features such as breathiness. We employ MEG as an experimental paradigm due to its superior precision to temporally (and spectrally) record cortical activity.

Previous human electrophysiological recording has shown that vowels themselves and signals that contain information like that of vowels (e.g., similar spectral envelope structure) can be categorized preconsciously and this is reflected in the evoked component amplitudes and latencies (Eulitz et al., 1995; Diesch and Luce, 1997; Poeppel et al., 1997; Diesch and Luce, 2000; Obleser et al., 2003a, 2004; Mizuochi et al., 2005; Mizuochi et al., 2007). With regards to the signals we employ in this experiment, we have several *a priori* conjectures. We hypothesize that each vowel category/transfer function should exhibit different amplitudes and latencies, especially for the N1/M100 component, as this component seems to index features

such as fundamental frequency, spectral envelope and signal complexity (Obleser et al., 2003a; Mäkelä et al., 2004; Obleser et al., 2006; Monahan et al., 2008). We also hypothesize that the more spectrally complex signals (those with more harmonics in the spectrum) will elicit faster M100 latencies and larger amplitudes (Diesch and Luce, 1997). Since the signals are relatively broadband (see *Methods*), we additionally hypothesize that M50 responses should be elicited, though we have no specifics concerning the direction of latency and amplitude modulations based on spectral structure (Chait et al., 2004; Howard and Poeppel, 2009). For the P2m, we hypothesize the more spectrally complex signals will elicit larger amplitudes in this evoked component (Shahin et al., 2005; Shahin et al., 2007). However, we are most curious as to how signals are processed within vowel categories; namely if there will be differences in the amplitude and latency of the evoked components recorded due to the harmonic structure of the source waveforms. The main thrust of the experiment is to explore the sensitivity of the major evoked components to the presence or absence of harmonics. Additionally, we aim to investigate the population-level resolution of auditory processing; we are curious to see if the timbral metamers also generate similar gross patterns of auditory processing. Ultimately then, the goal is to link low-level features and processing with the complex cortical parsing of auditory signals, with signals that may be processed using top-down (e.g., cognitive, phonetic) information (Sinnott et al., 1997).

Materials and Methods

Participants

Forty-six normal-hearing adult participants participated in non-invasive cortical magnetic field recording. Participants were compensated (\$10/hour) for their involvement. Presentation of stimuli and responses was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Before the start of each experiment, informed written consent was obtained from each participant. Participants were also asked to fill out a brief questionnaire documenting their musical or vocal training and experience.

Stimuli

Experimental signals were generated with MATLAB (v7.8 , R2009a, The Mathworks, Natick, MA). Seven complex source signals were filtered using the transfer functions for the American English Vowels /a/, /i/, and /u/ (Hillenbrand et al., 1995). Vowel bandwidth measurements were likewise derived from ecological token measurements (Fant, 1972). Source signals were generated using Fourier synthesis and the transfer functions were generated from an all-pole filter derived from measurements of the mean frequency values (Hz) of the first three formants for male speakers. Signal duration was 250 msec with eleven msec \cos^2 onset and offset ramps; this duration was chosen because it is (i) the approximate average length of a vowel and (ii) allows for robust elicitation of evoked signals in a passive listening condition. Signals were sampled at 44.1 kHz with 16-bit resolution. These signals are identical to those employed in a separate psychophysical and modeling experiment. Figure 1 illustrates the main properties of the synthesized signal formant (transfer function) structure. The peaks corresponding to the first three formants are

clearly visible, in addition to the spacing between formants and the spectral tilt for each vowel. For the /i/ and /u/ signals (high vowels), the proximity of the first formant to the fundamental frequency is readily apparent; for the /a/ and /u/ (back vowel signals), the close spacing of the first two formants is easily visible.

The experimental materials were designed to be (i) be ecologically valid, (ii) create timbral differences within and across signal classes and (iii) test the limits of the resolution of auditory perception (via timbral metamers). To that end, care was taken to select and synthesize source waveforms and filters (transfer functions) that satisfied these criteria. Since speech identification and processing is heavily reliant on timbre (Culling and Darwin, 1993; Fellowes et al., 1997; Gfeller et al., 1998; Dissard and Darwin, 2001), we chose to use as the basis for our experiment a source waveform that mimicked the glottal excitation pattern and modified its structure to induce timbral differences. Further ecological validity was added by filtering the source waveforms through the different vowel transfer functions as vowels are differentiated by their (i) timbre, (ii) vocal tract filtering characteristics and (iii) linguistic information. We chose vowel transfer functions that are well-separated (theoretically) timbrally and linguistically, namely the American English vowels /i/ (high front unrounded vowel), /u/ (high back rounded vowel) and /a/ (low back unrounded vowel). These vowels were chosen as they are the 'point vowels' and are well separated linguistically and perceptually. The source waveforms had a fundamental frequency (F0) of 150 Hz (gender-neutral fundamental voice frequency)

and contained a maximum of twenty-three harmonics to approximate the glottal excitation pattern (Cleveland, 1977; Miller, 1989).

Seven different source waveforms were employed to create a timbral continuum. The signals were as follows: sawtooth waveform approximating the glottal excitation pattern comprised of twenty-three harmonics, sawtooth waveforms with every sixth, fifth, fourth, third, and second harmonic removed and a square wave where the maximum harmonic is the twenty-third; the sawtooth wave with every second harmonic removed and the square wave were the timbral metamers. Each source waveform was filtered digitally, for a total pool of twelve signals. Signal power was equalized between signals (i.e., filtered source waveforms) via the root-mean-square (RMS) of the signal. Figure 2 illustrates the construction of several stimuli for the vowel /i/. This figure displays the temporal evolution of the source waveforms over several cycles, the effect of the transfer function on the source waveform and the power spectral density (PSD) estimate of the filtered source waveform spectral structure. Signals shown are the sawtooth wave approximating the glottal excitation pattern (top row), sawtooth wave with every fourth harmonic removed (middle row) and sawtooth wave with every second harmonic removed (bottom row). The effect of the harmonics present in the spectrum on source waveform structure (left column) to produce timbre differences is evident in the differences between the temporal structures of the signals. The oscillation of the waveform at F0 and the harmonics present is clearly in the middle column of the figure. The right column illustrates differences in spectral structure in the frequency

domain between the signals; presence and absence of source waveform harmonics is indicated by the ‘gaps’ – discontinuities introduced by the removal of harmonics. Figure 3 illustrates the timbral metamers for the /a/ signals. The left column illustrates the temporal structure of the sawtooth wave with every second harmonic removed (2F0 spacing) and the square wave source waveforms. There is a slight phase difference between the signals; the sawtooth wave begins at the minimum of the cycle, while the square wave begins at the maximum. Visualization of the spectral information (PSD) indicates that the signals are identical in the frequency domain and thus should be perceived as identical.

Experimental Procedure

Prior to the start of the experiment, two auditory pretests were performed. The first pretest used 250 and 1000 Hz sinusoidal signals of 400 msec duration presented at ~65 dB SPL; this pretest was used to establish that the participants generated an observable auditory-evoked response. On average, the 1000 Hz signal generated a more robust response. Selection of maximally responsive auditory channels was derived from a second pretest. The second pretest used 250, 1000 and 2000 Hz sinusoidal, square wave and sawtooth wave signals of 100 msec duration presented at ~65 dB SPL. This pretest was employed for two reasons: (i) to provide information about the relationship between signal frequency and waveform type (as this is somewhat underspecified (Roberts et al., 2000)) and (ii) provide information about the evoked response amplitude and latency distributions of the source waveforms used in the experiment. Five channels from source and sink for each

hemisphere (twenty channels total) with the maximum measured field deflection common to all waveforms and frequencies presented employed in the second pretest were used for analysis of the responses to the experimental materials. The ISI for the pretests ranged between 300 and 1000 msec; the analysis frame length was from 100 msec pretrigger to 400 msec posttrigger for the first pretest and from 115 msec pretrigger to 275 msec posttrigger for the second pretest.

For the main timbral experiment, participants were randomly assigned to one of two separate experimental groups: one group was required to passively listen to signals with the transfer functions corresponding to /a/ and /i/; the other /a/ and /u/. Several participants were rejected due to insufficient signal-to-noise ratios in each pool (/a/ - /i/: eight participants; /a/ - /u/: 9 participants). Several additional participants were rejected due to having a reversed auditory field topography for the M100 (/a/ - /i/: one participant; /a/ - /u/: two participants). The final number of participants in the /a/ - /i/ pool was fourteen (out of twenty-three); for the /a/ - /u/ pool eleven (out of twenty-one). Ten participants in the final pool for the /a/ - /i/ assignment stated musical experience, two participants stated they were currently practicing and two stated they had ear training. Ten participants in the final pool for the /a/ - /u/ assignment stated musical experience, two participants stated they were currently practicing and one stated they had ear training. The final pool of participants for both assignments consisted primarily of native English speakers, with three participants in the /a/ - /i/ assignment being bilinguals for Chinese, Russian and Spanish natively. Two participants in the /a/ - /u/ assignment were native speakers of

Czech and Korean. Each signal was presented 130 times, pseudorandomly interleaved, were passively attended to and presented at ~65 dB SPL. To maintain participant vigilance, a distracter task was incorporated. Approximately Gaussian white noise (400 msec duration) was used as a target during the experiment and was pseudorandomly presented with the signals (~ 17% of total). Participants had to press a button in response to the noise target; these trials were excluded from analysis.

Delivery

All experimental stimuli were presented using a Dell Optiplex computer with a M-Audio Audiophile 2496 sound card (Avid Technology, Inc., Irwindale, CA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Stimuli were delivered to the participants binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). The inter-stimulus interval varied pseudo-randomly between 700 and 1100 msec.

Recording

Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was DC-200 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. Data were noise reduced using time-shifted principal component analysis (de Cheveigné and Simon, 2007), the trials averaged and baseline corrected offline (artifact rejection ± 2.5 pT). Signals for the evoked response (i.e. after trial averaging) were low-pass filtered using a fourth order elliptical filter with a cutoff frequency of 40 Hz, 0.5 dB

peak-to-peak ripple and at least 60 dB stopband attenuation. Time-shifted principal component analysis does not remove eye-blink artifacts, but trials with excessive frontal field deflections were excluded from the analysis.

Evoked Response Analysis

The magnetic field evoked response data for the main timbral experiment were analyzed so as to (i) take advantage of the ability of MEG to capture-within participant responses and (ii) account for the univariate design of the experiment.

Peak RMS and latency values for M50 (search window: 35 to 84 msec), M100 (search window: 85 to 170 msec) and P2m (search window: 171 to 250) responses were determined using the localizer pretest channels. The signal evaluation window ranged from 250 msec pretrigger to 600 msec posttrigger. Evoked response peak RMS and latency values for data channels selected from the second pretest were averaged across participants and evoked responses were plotted topographically to confirm the response. In addition, when a participant's data did not show an auditory cortex magnetic field topography for a given condition, after averaging and filtering, the data (peak RMS and latency) were excluded from further analysis (Luo & Poeppel 2007). Significance of the peak RMS and latency values collected was evaluated using General Linear Models (GLMs) (see below).

Due to the univariate design of the experiment, RMS vectors for each hemisphere and condition for each participant were collected and placed into matrices

for further analysis. Grand average RMS measurements (taken from filtered and baseline corrected data) were computed via RMS of the RMS for each vowel transfer function and condition, by hemisphere. Differences between the grand average responses for amplitude were calculated via bootstrapped 95% confidence intervals (CIs) -- see below (Chait et al., 2010).

Statistical Analysis of Responses

Statistical analyses of the evoked response data were analyzed using R 2.10.1. GLMs were implemented using the “languageR” statistical package (Baayen, 2010). For the experimental pretest, GLMs were implemented for the M100 peak RMS and latency for the factors Hemisphere, Frequency, and Waveform with Participant as the variable interaction. The factors and interactions for the pretest GLMs were (i) Hemisphere x Frequency x Waveform, (ii) Frequency x Waveform and (iii) Waveform and (iv) each factor independently. The GLMs for each main experimental assignment evaluated M50, M100 and P2m peak RMS and latency for the factors Hemisphere, Vowel, Waveform and Harmonic Structure, with Participant as the variable interaction. The initial models considered the following factors and interactions: (i) Hemisphere x Vowel x Waveform x Structure, (ii) Vowel x Waveform x Structure, (iii) Waveform x Structure, (iv) Structure and (v) each factor independently.

Additional statistical analyses were implemented separately for participants by dividing the participant pool into two different groups, those that indicated they had

musical training and those without. For the participants who indicated they had musical training, the factors were Training Type (formal or informal), Years playing the instrument, whether or not the instrument was currently being played, and whether or not the participant had ear training. The factor interactions for the analysis of the participants with musical training were (i) Training Type x Years Played x Current Playing x Ear Training, (ii) Training Type x Years Played x Current Playing, (iii) Training Type x Years Played and (iv) Training Type. A final statistical analysis for each experimental assignment compared the responses between those with musical training and those without (Crummer et al., 1994; Lakatos, 2000; Samson, 2003; Chartrand and Belin, 2006). Significance for all statistical evaluations was performed at the $\alpha = 0.05$ level.

Amplitude differences were evaluated via 95% Bca bootstrap confidence intervals on the difference between RMS vectors from the hemispheres for (i) each vowel overall (ii) each condition by hemisphere and (iii) the timbral metamers via MATLAB's Statistics Toolbox. The procedure was as follows: for each condition (for vowels, all conditions) the individual RMS vectors from each participant were collected into a matrix, with each participant vector in a single row. The difference was then calculated via subtraction (e.g., LH – RH, /a/ - /i/) of one matrix from the other. Five consecutive time points where either the lower or upper bounds were both less than or greater than zero were taken as significant (Chait et al., 2010). For CI calculation, there were 500 bootstrap resamples.

To account for individual cortical differences in the processing of the experimental signals, Bca 95% bootstrap confidence intervals were calculated for the (i) correlation between the participant latency values (for each hemisphere and condition) and timbral complexity measures using Spearman’s rank correlation and (ii) participant differences in latency values for each hemisphere and condition using the “boot” package in R 2.10.1 with 1000 resamplings (Davison and Hinkley, 1997; Canty and Ripley, 2010). Since the signals employed were completely steady-state, their timbre thus depended entirely on frequency content. In addition, since they were ecologically based, we decided to use a measure that reflected this fact and have been implicated in the electrophysiological literature as possibly affecting evoked response activity. The correlations were between the evoked component latencies and the spectral center of gravity. Spectral center of gravity is an amplitude-weighted measure that indicates where energy is concentrated in the frequency domain (i.e., a measure of central tendency). Amplitude-weighted spectral information was derived via power spectral density (PSD) estimates of the digitized, RMS-equalized waveforms presented to the experimental participants using Welch’s method using MATLAB’s Signal Processing Toolbox. The PSD was estimated only for the harmonics present in the spectrum. Spectral center of gravity was calculated as follows:

$$SCG = (\sum_n n \times A_n) / \sum_n A_n$$

where n denotes the frequency in Hz, and A is the amplitude in decibels (dB). We also included a version of the SCG where frequency is log-transformed (log base 10) to improve ecological validity; a log-log representation of the signals would more

accurately represent how the auditory system parses the signal (Hartmann, 1998; Moore, 2004).

Results

M100 Amplitude and Latency Pretest Results

The main objective of the pretests was to determine the channels that recorded from the auditory areas responding most robustly to the signals presented; a minor objective was to determine response properties to signals that would be decomposed harmonically (sawtooth and square waves) as in the main experimental signals, in addition to signals that have been traditionally used to determine basic auditory function (sinusoidal waves). Data from two participants from the final pool of participants were excluded due to not generating robust SNRs for a majority of pretest conditions. Figures 4-6 summarize the data from the second pretest, analyzed according to hemisphere, waveform and oscillation frequency. Figure 4 illustrates the normal approximations to the M100 amplitudes and latencies for each hemisphere. Though not apparent in the boxplots, the data replicate several previous findings concerning M100 amplitude and latency responses. First, the signals elicited typical M100 responses; namely that the RH exhibited a greater M100 amplitude and faster M100 latency; this was verified via statistical analysis. Inter-participant variability was also high; aside from a few exceptions, there was no difference in either M100 amplitude or latency between waveform oscillation frequencies.

Figure 5 illustrates the mean M100 amplitude values, separated by hemisphere and identified according to frequency (250, 1000, 2000 Hz) and waveform (sawtooth, square, sinusoid), while Figure 6 illustrates the mean M100 latency data, with the same separations as in Figure 5. This data visualization makes the relationships observed to be statistically significant more readily apparent. Statistical evaluation of M100 latency data for each of the factors assessed independently (Hemisphere, Frequency, Waveform) found significant predictors to be each hemisphere (coef = -3.211, SE = 1.195, $t = -2.690$), with the RH exhibiting a faster M100 latency. No significant differences were observed for frequency and waveform assessed independently. Based on the observed difference between hemispheres we performed *post hoc* tests, evaluating waveform structure and oscillation frequency independently as well as their interaction with each other, within each hemisphere. Statistical evaluation of LH M100 latency found 2000 Hz signals to be significant predictors compared to the 250 and 1000 Hz signals (coef = 5.793, SE = 2.083, $t = 2.780$). No significant differences were observed between waveform types in the LH. Assessment of the interaction Frequency x Waveform, 2000 Hz signals were again found to be significant predictors (coef = 11.788, SE = 3.669, $t = 3.210$) as well as 2000 Hz sinusoidal wave signals (-15.255, SE = 5.093, $t = -3.000$). Evaluation of the RH M100 responses found no significant differences for oscillation frequency and waveform evaluated independently; however, the interaction Frequency x Waveform for 2000 Hz sinusoidal signals was observed to be a significant response predictor (coef = -7.649, SE = 4.032, $t = -1.900$).

Statistical analysis of M100 amplitude data for each of the factors assessed independently found Hemisphere to be a significant response predictor (coef = 18.271, SE = 2.980, $t = 6.133$) as well as the square wave signals (coef = -5.735, SE = 3.711, $t = -1.515$); both the RH and the square wave signals exhibited larger M100 amplitudes. Based on these results we performed *post hoc* tests in the same manner as the tests on the M100 latency data. Evaluation of the LH responses with each factor assessed independently found 1000 Hz signals to be significant predictors (coef = -12.074, SE = 4.113, $t = -2.936$) in addition to square waveform signals (coef = 9.819, SE = 3.988, $t = 2.462$). Assessment of the interaction between oscillation frequency and waveform found significant response predictors to be square wave signals overall (coef = 23.292, SE = 6.776, $t = 3.438$) and 1000 Hz square wave signals (-23.724, SE = 9.393, $t = -2.526$).

Statistical analysis of the three-way interaction latency data for the interactions Hemisphere x Frequency x Waveform found 2000 Hz signals (coef = 13.198, SE = 3.617, $t = 3.650$), specifically 2000 Hz signals in the RH (coef = -12.542, SE = 4.971, $t = 2.520$) and 2000 Hz sine waves (coef = -16.633, SE = 5.023, $t = -3.310$) to be significant response predictors. For the interaction Frequency x Waveform, 2000 Hz signals (coef = 6.356, SE = 2.556, $t = 2.490$), sine waves (coef = 5.184, SE = 2.555, $t = 2.030$) and most specifically 2000 Hz sine wave signals (coef = -12.248, SE = 3.552, $t = -3.450$) were observed to be significant response predictors. Analysis of the three-way interaction of amplitude data for the interaction Hemisphere x Frequency x Waveform found square wave signals (coef = 22.394, SE

= 8.982, $t = 2.493$) to be significant predictors of response amplitude. Analysis of the interaction Frequency x Waveform found square waves to be significant predictors (coef = 19.906, SE = 6.707, $t = 2.968$) and specifically 1000 Hz square wave signals (coef = -13.310, SE = 9.261, $t = -2.265$).

Analysis of the sensors that responded most robustly revealed that across conditions and within participants, the same sensors tended to respond to the signals presented. Analysis of the response density distribution revealed that participant responses were highly variable, but that within-participant neuronal processing was stable.

Preattentive Timbre Classification Results

The majority of subjects did not exhibit robust M50 or P2m responses, with the M50 being generated more often than the P2m. As such, we decided to focus our initial analysis on the M100 responses generated to the signals presented, which were reliably generated. One participant, in the /a/ - /i/ pool, had a very left-lateralized M100 response. For the statistical analyses, where a participant did not exhibit an evoked response (via evaluation of the magnetic field topography), missing values were *not* imputed with the series mean to stay as close as possible to the observed data. Mean values from both hemispheres were summarized according to factors previously reported to affect to be affecting M100 latency in the literature (hemispheric differences, vowel identity) as well as the factors which we were specifically investigating. Histograms and smooth density estimates of the M100

latency data were skewed left slightly (though the latency data in the /a/ - /u/ assignment were more normally distributed), while the peak RMS response data were skewed right, a result typical of response power values.

/a/ - /i/ Experimental Assignment

Analysis of the M100 latency and peak RMS data across participants (averaging of individual evoked response latency and peak RMS values) for the /a/ - /i/ experimental assignment by hemisphere and vowel revealed that the data replicated several previous findings. First, there was an asymmetry between the hemispheres: the RH displayed a faster latency and larger amplitude than the LH. Also replicating previous data, there was no overall consistent difference between the evoked responses to each vowel signal class (Mizuochi et al., 2007).

Overall, Hemisphere was observed to be a consistent predictor of M100 amplitude and latency, with the RH exhibiting greater amplitude and faster latency values than the LH. The only other consistent predictor was Waveform. The observation that there was no significant difference between /a/ and /i/ signals overall, in latency and amplitude is in line with previously reported data concerning signals of a vocal nature (Mizuochi et al., 2007). Based on the observations from the initial statistical tests, we conducted *post hoc* tests where the data from each hemisphere was separated and the remaining factors were evaluated independently and interactively.

Statistical analysis via GLMs found Hemisphere (LH vs RH) and Waveform (sawtooth vs square) to be consistent predictors of M100 amplitude overall. RH amplitude (coef: 14.637, SE = 3.752, $t = 3.901$) and latency for the RH (coef = -2.536, SE = 0.956, $t = -2.650$) and square source waveform (coef = -3.691, SE = 1.290, $t = -2.860$) were significant predictors when evaluated independently. Significant results were obtained for Hemisphere when evaluated both interactively for amplitude (RH minimum significance: coef = 15.842, SE = 5.716, $t = 2.772$) and latency (RH minimum significance: coef = -2.777, SE = 1.329, $t = -2.090$). No specific SCG value was observed to be a significant response predictor for either amplitude or latency. *Post hoc* tests where we separated the data from each hemisphere and then evaluated the remaining factors independently and interactively found several significant predictors. For the LH M100 amplitude responses, significant predictors were observed to be the /i/ signals (coef = 6.155, SE = 2.586, $t = 2.380$) and the square waveform signals (coef = 14.919, SE = 4.536, $t = 3.289$) for the interaction Vowel x Waveform; the square waveform signals were also observed to be a significant predictor when Waveform was evaluated independently (coef = 10.261, SE = 3.435, $t = 2.987$). For the RH data, the square waveform signals were observed to be a significant predictor with Waveform evaluated as a factor independently (coef = 13.030, SE = 4.100, $t = 3.179$). For the M100 latency, square waveform signals in the RH were observed to be significant response predictors when evaluated independently (coef = -4.729, SE = 1.868, $t = -2.530$). Within vowel signals, the RH was observed to be a significant response predictor for the /a/ signals for both M100 amplitude (minimum significance coef = 16.670, SE = 5.551, $t = 3.003$) and latency

(minimum significance: coef = -2.874, SE = 1.308, t = -2.200). For the /i/ signals, the only significant predictor was the RH for M100 amplitude (coef = 14.790, SE = 5.315, t = 2.783).

Figures 7-9 summarize the analysis of bootstrapped confidence intervals and grand averaged evoked responses. Figure 7 illustrates the grand averaged RMS waveforms (RMS of RMS) and the time points observed to be significant for the RMS amplitude via bootstrap resampling, separated by hemisphere for several conditions. One striking characteristic of the data is that across conditions, there is an asymmetry in the peaks of the evoked response between the hemispheres. In the LH, there is a distinct peak in the time range where a M50 response would be expected to be observed (~90 msec) in addition to the M100 peak. As previously mentioned, not all participants showed a M50; however, for those participants who exhibited a robust M50 response, the response (i.e. magnetic field topography and RMS peak) was very left-lateralized. We limited our analyses of the bootstrapped RMS amplitude differences to the time windows of interest for the evoked responses based on our initial hypotheses and the previous literature. Due to the large variance observed, we decided to compute grand averages of the magnetic field deflections of the filtered and baseline corrected participant data to determine if the responses were likely generated in auditory cortex based on the topography observed.

There was no consistent relationship between conditions regarding the time points where the differences between the response amplitudes generated by the /a/

and /i/ signals. Several time points in the sustained field (~300 msec post-onset) were observed to contain significant RMS amplitude differences, but this observation was not consistent. Comparison with the grand averaged magnetic field deflections indicated that the sustained field time ranges did display a sink-source distribution similar to that of the M100, but its relationship to the M100 and its generators is unclear.

Figure 8 illustrates the grand averaged RMS responses for each of the metamers in each vowel class. The RMS responses parallel each other closely; in fact, they almost completely overlap each other in the temporal domain. Bootstrap analyses of the metamers for each vowel signal found significant differences in amplitude in the LH prior to M100 generation and in the sustained field. However, it is not clear if these differences between metamers are significant, due to the large variance in significant time points observed for all signals across both classes. Figure 9 illustrates the differences between each vowel signal (all conditions aggregated), separated by hemisphere. In parallel with the analyses of means of the overall responses, especially for the M100, there are very few differences between the vowels; the RMS time courses are essentially identical. Figure 10 illustrates the relationship between the RMS grand averages, magnetic field grand averages and grand average topography. The top panel illustrates the temporal evolution of the evoked responses; data are taken from 28 sensors common to all participants in both hemispheres. Three distinct peaks can be observed one corresponding to a typical M50 response (~90 msec), another corresponding to the M100 (~150 msec) and a

third in the sustained field (~308 msec). The topographies for each of the responses are displayed below the grand averaged magnetic field deflections. The M50 topography indicates that it has an opposite dipolar pattern to that of the M100 and sustained field and is extremely left-lateralized.

Based on the asymmetries observed in the RMS grand averages, we followed our initial analyses with *post hoc* tests on the M50 responses. We analyzed data from each hemisphere separately, due to the M50 responses being extremely left-lateralized and previous data indicating that the LH may play a special role in distinguishing between variations in source waveform spectral structure (Tiitinen et al., 2005). Statistical evaluations of the M50 amplitude and latency only found significant differences between the hemispheres for both amplitude (coef = -20.093, SE = 3.465, t = -5.799) and latency (GLM: coef = -5.462, SE = 1.570, t = -3.480); no other significant differences within hemisphere for any other factors or interactions between factors was observed.

Analysis of the bootstrap distributions and quartile plots using SCG as a proxy for harmonic and vowel structure indicated that within-participant responses to the signals were consistent. Analysis of the bootstrapped latency differences revealed that the across-participant responses were highly variable and that bootstrap estimations and distributions were not appropriate for assessing the differences.

/a/ - /u/ Experimental Assignment

Figure 11 illustrates mean M100 amplitude and latency for each hemisphere and vowel. In replication of previous data, the RH exhibits a greater M100 amplitude and shorter latency. Also in agreement with previously reported data, the /a/ signals exhibit a greater M100 amplitude than the /u/ signals, as well as a shorter M100 latency (Poeppel et al., 1997; Mizuochi et al., 2007). Hemisphere, Vowel and SCG were observed to be consistently significant predictors of M100 amplitude and latency overall. SCG as a predictor in the /a/ - /u/ assignment differed from SCG in the /a/ - /i/ assignment in that specific SCG values were identified as being specific predictors using GLMs. As with the /a/ - /i/ assignment, we separated the data from each hemisphere and evaluated the remaining factors independently and interactively.

For analysis of the M100 amplitude data using GLMs, the RH (minimum significance: coef = 21.448, SE = 3.982, $t = 5.386$) and the /u/ signals (minimum significance: coef = -8.869, SE = 3.999, $t = -2.217$) were consistently found to be significant response predictors; the same factors were also consistent predictors for the M100 latency (RH minimum significance: coef = -4.150, SE = 1.380, $t = -3.010$; /u/ minimum significance: coef = 3.547, SE = 1.386, $t = 2.560$). Additionally, there were several SCG values (linear and log frequency) found to be significant response predictors for M100 amplitude and latency, but significance was not consistent across models. Incidentally, the SCG values that were found to be significant predictors came from /a/ signal class. As done previously, we performed *post hoc* tests separating the hemispheres and evaluating the remaining factors independently and

interactively. For the LH amplitude responses, the /u/ signals were found to be significant predictors for the interaction Vowel x Waveform (coef = -8.821, SE = 3.021, $t = -2.920$) and when evaluated independently (coef = -9.012, SE = 2.808, $t = -3.209$). For the RH amplitude responses the /u/ signals (coef = -15.015, SE = 3.173, $t = -4.733$) and the square waveform signals (coef = 13.407, SE = 6.120, $t = 2.191$) were found to be significant predictors for the interaction Vowel x Waveform; the /u/ signals were also found to be significant predictors when evaluated independently (coef = -16.233, SE = 2.971, $t = -5.464$). The linear SCG values that were found to be significant responses for M100 amplitude were all the in the RH and corresponded to the sawtooth wave with every second harmonic removed (coef = 34.398, SE = 7.223, $t = 4.762$), the sawtooth wave with every fourth harmonic removed (coef = 16.099, SE = 7.223, $t = 2.229$) and the square wave (coef = 26.095, SE = 7.440, $t = 3.616$) source waveforms. The log SCG values that were found to be significant response predictors for M100 amplitude were likewise contained in the RH and were the /a/ timbral metamers (a-2-rem: coef = 24.832, SE = 7.366, $t = 3.371$; a-sq: coef = 32.326, SE = 7.151, $t = 4.520$).

The /u/ signals were observed to be significant response predictors of M100 latency in both hemispheres (LH minimum significance: coef = 2.890, SE = 1.317, $t = 2.190$; RH minimum significance: coef = 4.417, SE = 0.878, $t = 5.030$). Linear SCG values observed to be significant response predictors (all /a/ signals) of LH M100 latency were the sawtooth wave with every third harmonic removed (coef = -7.333, SE = 3.237, $t = -2.270$), second harmonic removed (coef = -6.841, SE = 3.142, $t = -$

2.180) and square wave (coef = -9.310, SE = 3.141, t = -2.960) source waveforms. In the RH, every /a/ waveform except for the sawtooth wave with every fifth harmonic removed was a significant predictor for their linear SCG values (minimum significance: coef = -5.959, SE = 2.091, t = -2.850); for the log SCG values, the sawtooth wave approximating the glottal excitation pattern (coef = -7.400, SE = 2.070, t = -3.580) and the square wave source waveforms (coef = -6.266, SE = 2.132, t = -2.940) were observed to be significant predictors. Within-vowel analyses of the /a/ signals revealed that the RH was a significant response predictor for both M100 amplitude and latency (amplitude minimum significance: coef = 21.625, SE = 4.440, t = 4.871; latency minimum significance: coef = -4.105, SE = 1.313, t = -3.130); the same was true for the /u/ signals (amplitude minimum significance: coef = 15.172, SE = 3.498, t = 4.337; latency minimum significance: coef = -3.871, SE = 1.353, t = -2.860). The only other significant response predictors were the sawtooth wave with every second harmonic removed for the /a/ signals (coef = 9.951, SE = 8.8034, t = 2.468) for the factor Harmonic and for linear SCG value (coef = 17.082, SE = 8.149, t = 2.096) for M100 amplitude.

Figure 12 illustrates the grand averaged RMS waveforms (RMS of RMS) and the time points observed to be significant for the RMS amplitude via bootstrap resampling, separated by hemisphere for the /a/ - /u/ metamers. As with the /a/ - /i/ signals, there was considerable variance in the time points found to be significant; the identical procedure was used to evaluate the RMS responses. As in the /a/ - /i/ assignment, there was an asymmetry between the hemispheres, with a distinct peak in

the time range expected for a M50 response occurring more often and robustly in the LH. There was no consistent relationship between conditions for the time points where the RMS amplitudes were observed to be different via bootstrap resampling. Again, there were several time points in the sustained field with a distribution similar to that of the M100, the relationship between the M100 responses generated and the sustained fields from this experiment assignment are unclear. Amplitude differences between the metamers were observed, but due to the large variance observed for all conditions, the significance is once again unclear.

Figure 13 illustrates the grand average magnetic field deflections and topographies for each vowel class across condition for the /a/ - /u/ assignment. Conventions used are the same as that of Figure 10. Visualization of the evoked response in the temporal domain reveals that the /a/ signals exhibited a greater M100 amplitude and shorter latency than the /u/ signals. Topographical illustration of the M50 and M100 evoked response distributions reveal that the responses have opposite sink-source distributions, indicating different patterns of neuronal activity. In the experimental pool the overall M50 response was not as left-lateralized for the /a/ signals.

Statistical analysis of the M50 responses found only significant differences between hemispheres for amplitude (coef = -5.071, SE = 2.077, $t = -2.441$) and latency (coef = -5.219, SE = 1.533, $t = -3.400$); as before, this may have been a consequence of the M50 being more reliably and robustly generated in the LH.

Effects of Musical Training and Native Language

For both experimental assignments, no statistically significant differences were observed in evoked response amplitude or latency for (i) musical/vocal training, (ii) musical or vocal training type (formal or informal), (iii) whether or not the instrument trained on was currently being played and (iv) native language. The negative result for musical training may have been due to several factors: the large number of subjects in the final pools with musical/vocal training, the stability of evoked response for the signal employed or perhaps the lack of a psychophysical task. The negative result for the effect of native language may have been due to the participants being in an American English immersive environment as well as the use of tokens that did not contain features of tonal language vowels (e.g., low-high tonal transitions).

Discussion

The data presented in this experiment reproduce several findings regarding the processing of auditory signals having formant or formant-like spectral structures. First, we reproduced the hemispheric asymmetries observed for the M100, with the RH exhibiting faster latencies and larger amplitudes than the LH. We also replicated previous findings for hemispheric asymmetry for the M50, with a LH bias in the generation of the M50 (Howard and Poeppel, 2009). We also replicated data from several studies regarding differences in vowel processing, namely that the evoked responses between the /a/ and /u/ signals differed in M100 amplitude, latency and

temporal evolution, while those between the /a/ and /i/ signals were essentially identical (Poeppl et al., 1997; Mizuochi et al., 2007). While our specific hypotheses regarding how harmonic structure affected evoked response latency and amplitude were not supported by the data, the hypotheses concerning vowel category processing were supported. Though the signals we employed in this experiment lacked several features of real vowel tokens (e.g., formant transitions, breathiness, pitch contour) their composition reflected crucial features of vowel formant structure and these features were reflected at a gross level in cortical processing.

The data recorded support the view that signals, especially vowels or vowel-like signals, can be categorized preattentively/preconsciously (Diesch and Luce, 2000; Vihla and Salmelin, 2003; Jacobsen et al., 2004a), at least by the time the M100 is generated (Gage and Roberts, 2000). Contribution of F0 to categorization and the evoked responses was minimal at best, as the signals all had the exact same fundamental frequency. However, the M100 responses observed occur rather late (~130-140 msec post-onset) in a range typical of signals with a fundamental of ~150 Hz (Roberts et al., 2000). The evoked response evolution was most likely due to the spectral envelope of the signals, which imposes specific dynamics on the temporal aspects of the signal in addition to shaping the frequency content. Spectral envelope has been identified as an especially salient dimension and this saliency (though not measured through psychophysics) seems to be reflected in the evoked response data. Statistical analyses as well as investigation of the grand averaged RMS and magnetic field deflection demonstrate that the effects observed were reliable and robust across

participants, though there was considerable variation across participants. However, the extreme degree of similarity between the /a/ and /i/ signals is not explainable by the data. These signals belong to well-separated vowel classes and thus have markedly different spectral envelope structures. The similarity between the signals may be due to 'averaging' of the formant values; if this theoretical averaging does take place, then the average of the first two formant values for /a/ may be close to the F1 value for that of /i/. This is linguistically plausible as the F1/F2 ratio is very salient to an observer and F1 value can aid in tasks such as vowel identification (Darwin et al., 1989; Sinnott et al., 1997; Vihla and Salmelin, 2003; Jacobsen et al., 2004a). Though we have no psychophysical data to support this notion, it is possible that phonetic information may have also contributed to the analysis of the signals employed. Previous MEG and psychophysical data has shown that, depending on context, phonological cues as well as basic sensory processes contribute to signal identification (Sinnott et al., 1997; Monahan et al., 2008). Once again, this is reflected in our replication of data concerning evoked responses to vowel and vowel-like sounds. It stands to reason then, that signal parsing by the auditory system as well as top-down information contributes to preattentive signal classification, though the specifics of this process still elude us.

Previous data has shown that the hemispheres are differentially sensitive to different aspects of the vowel and vowel-like signals (Tiitinen et al., 2005). Specifically, variations in harmonic structure are processed preferentially in the LH, while spectral envelope is preferentially processed in the RH. Our results in the /a/ -

/u/ experimental assignment mirror that of Tiitinen et al (2005), where the excitation produced by /a/ and /u/ signals using varying source waveforms was explored using passively attended stimuli. The fact that our results do not agree completely may be due to the fact that particular study used a more varying set of source waveforms (glottal excitation, sine wave, aperiodic noise) than we employed; all our source signals were harmonic in nature. Though we created a timbral continuum, the majority of source waveforms were sawtooth in some respect. Another possible cause of the discrepancy is that there was a wide amount of variation between participant evoked responses to the harmonic structure of the source waveforms, though responses within categories were consistent. Due to the variance observed in the data, it appears that neuronal computation within signal categories varies greatly. Psychophysical tasks or participant training could enhance perception of subtle timbral differences. For example, evoked responses can be enhanced (faster response latencies, greater amplitudes or both) if participants are trained (musical or otherwise) to detect specific signals, signal features or discriminate between signals (Crummer et al., 1994; Hirata et al., 1999; Brattico et al., 2003; Chartrand and Belin, 2006). Though we did observe some statistically significant responses to several of the source waveforms, the degree of variation observed within categories precludes us from making any definitive statements about that significance and further investigation is warranted.

While this study was not concerned with the cortical generators of the evoked response, the evoked response temporal evolution supports several previous

observations regarding the generators of the responses involved. It is well-established that the M100 has several generators that contribute to its slope and latency (Eulitz et al., 1995; Pantev et al., 1995; Lütkenhöner and Steinsträter, 1998; Lütkenhöner, 2003; Okamoto et al., 2008; Howard and Poeppel, 2009). Relative to the M50, the M100 is much larger and has a tendency to overlap somewhat with the M50. In addition to the M50 asymmetries observed, examination of the grand averaged RMS and magnetic field deflection topography illustrate this overlap and dominance of the M100 relative to the M50, indicating that the generators of the responses observed in this experiment are most likely identical to those found via dipole analyses (Godey et al., 2001; Yvert et al., 2001; Yvert et al., 2005; Inui et al., 2006; Howard and Poeppel, 2009). Observation of M50 responses may have been aided by fact the signals employed were relatively broadband in nature (Chait et al., 2004). The generators of the sustained field deflections as well as their functional significance is unclear at this point in time; their topography indicates that they share many of the same generators as the M100, but this does not lend itself to an explanation of what the sustained field is indexing.

The experimental paradigm and materials reported could be expanded in several ways. Most obviously, a greater number of transfer functions corresponding to different vowels (e.g., vowels in the middle of the vowel space) could be employed. The transfer functions could also be expanded to include data derived from the measurements of female and child vocal measurements; the M100 especially has been shown to be sensitive to the formant structure and has exhibited different

response profiles to the spectral structure of vowels taken from males, females and children; the different genders and ages have different vocal tract lengths (Fant, 1972) which are directly related to the formant structure. While we aimed to but not find any differences in the evoked response due to source waveform structure, it would be useful to expand the source waveforms to those such as used by Tiitinen (2005). It may well be the case that harmonic waveforms, regardless of whether or not they resemble that of the glottal excitation pattern, may be preferred relative to other waveforms (e.g., sine waves, aperiodic noise).

Overall, we find that spectral envelope information allows synthesized signals to be categorized preattentively, as indexed primarily by the M100 evoked response component. However, differences in the latency and amplitude of this component may lack the resolution to distinguish between certain signals (e.g., /a/ and /i/). The data support the view that this preattentive categorization is derived from spectral envelope information, with possible contributions from top-down information and the fundamental frequency of the signal. The data do not support the view that differences in harmonic structure are reflected in the evoked response; however, this may be due to the observers lacking specific instruction concerning the nature of the signals employed as well the lack of psychophysical measures or training in the experiment (Pastore et al., 1990).

Figures

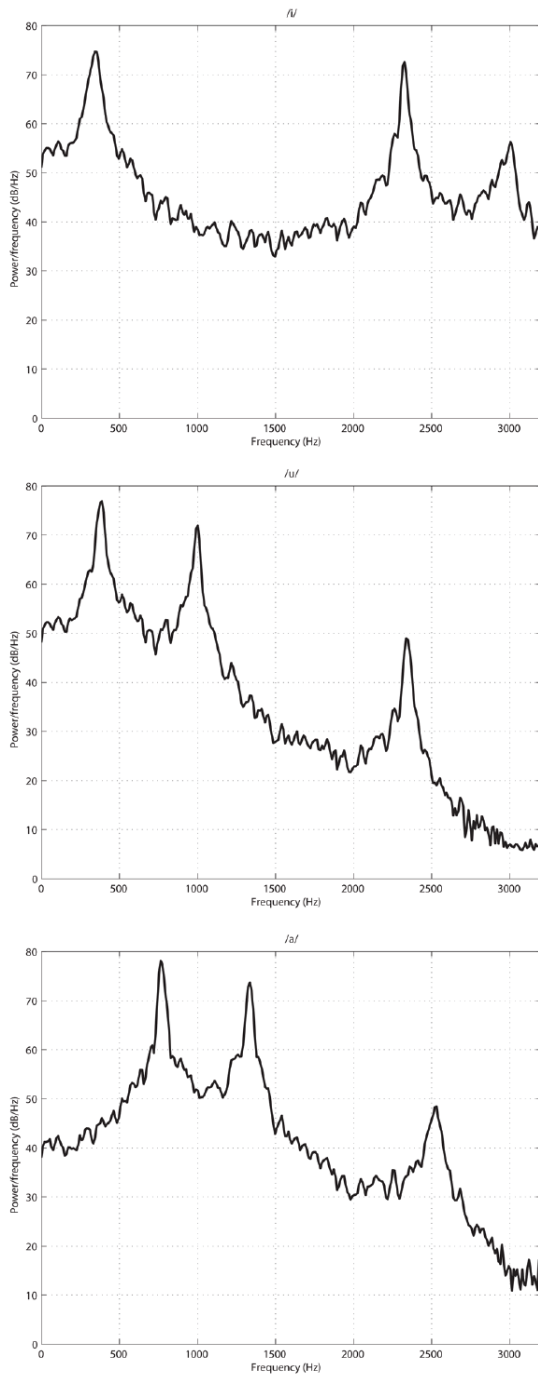


Figure 1. Spectral envelopes of vowel transfer functions employed in the experiment.

Top panel illustrates the transfer function for /i/, middle panel transfer function for /u/ and the bottom panel the transfer function for /a/ (/a/). The x-axis is frequency (Hz);

y-axis is the power spectral density (PSD) estimate (dB/Hz). Power spectral density estimates were computed using Welch's method. The vowels /i/ and /u/ are the high vowels and /a/ and /u/ are the back vowels. The transfer functions are acting as filters on a white noise signal. Transfer functions were created by taking data derived from American English male formant values (Hz) for each vowel (Hillenbrand et al., 1995) and converting the values into the coefficients of a transfer function to filter the various source waveforms employed. The visualization illustrates several important features of each vowel. For the high vowels, it is easy to see that the value of the first formant is relatively close to the fundamental frequency (150 Hz) and that the spectral valley below the first formant is shallow compared to the transfer function for /a/. For the back vowels, the first and second formant vowels are rather close to each other and the spectral tilt is more pronounced than in the front vowel /i/.

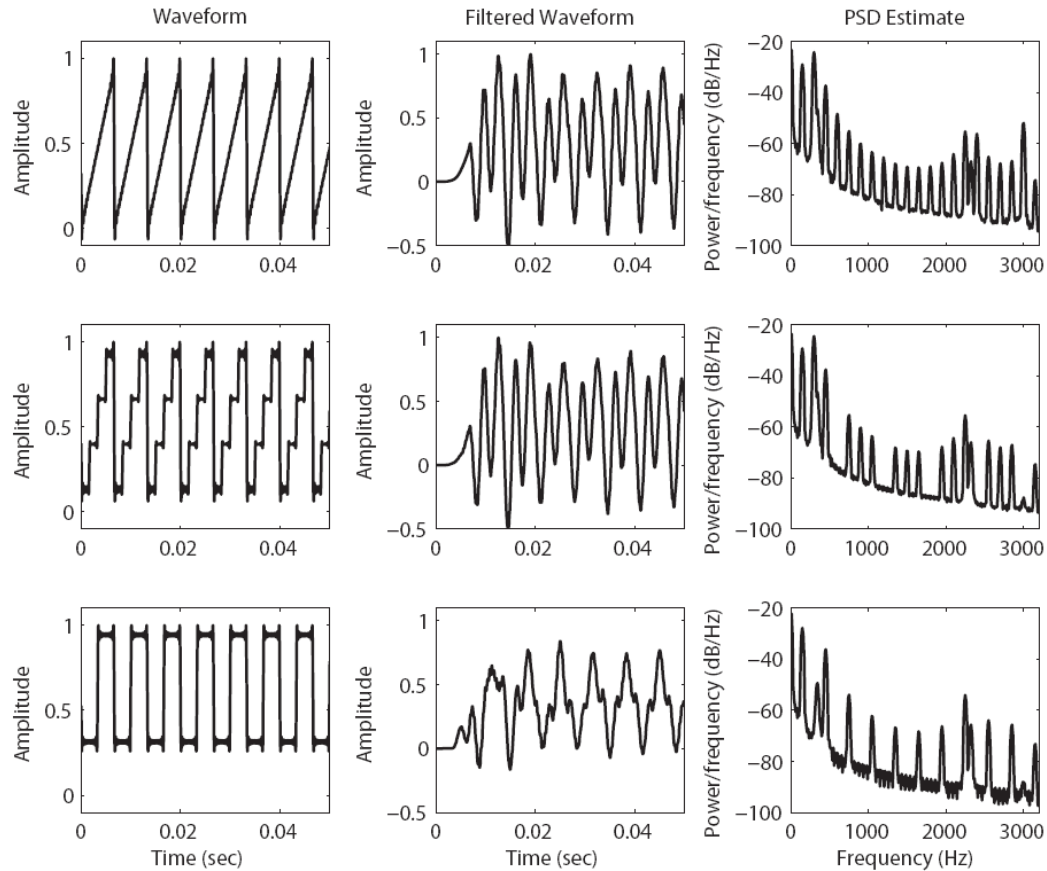


Figure 2. Illustration of stimuli construction, with transfer functions corresponding to the vowel /i/ as an example. Left column illustrates the structures of several source waveforms in the temporal domain; middle column illustrates the structure of the source waveform after filtering by the transfer function; right column illustrates the PSD estimate of the filtered waveform. Top row illustrates the sawtooth wave approximating the glottal excitation pattern; middle row the sawtooth wave with every fourth harmonic removed; bottom row sawtooth wave with every second harmonic removed. Source waveforms were constructed using Fourier synthesis. Visualization of source waveform structure in the temporal domain illustrates (i) the effect of Fourier synthesis in producing the harmonic waveforms and (ii) the effect of

selectively removing harmonics to produce discontinuities (in the frequency domain) and timbre differences between the source waveforms. Visualization of the transfer function (filters) on the source waveforms illustrates the steady-state temporal dynamics of the signals. The steady-state nature of the signals is readily apparent, i.e., the waveforms oscillate at a combination of the fundamental frequency and the harmonics present. Though somewhat difficult to see, close inspection of the sawtooth wave approximating the glottal excitation pattern and the sawtooth wave with every fourth harmonic removed reveals that (i) the essential sawtooth nature of the signals is preserved, despite differences in the harmonic structure and (ii) the selective removal of harmonics results in slight differences in the temporal evolution of the signals. PSD estimates more clearly reveal the timbral differences in frequency domain. The spectral peaks that stand out most prominently are at the values of the formants; though synthetic the signals employed exhibit the comb structure typical of vowel tokens and vowel-like signals. The effect of selectively removing harmonics is also clear; there are ‘gaps’ in the spectrum where those harmonics are not present.

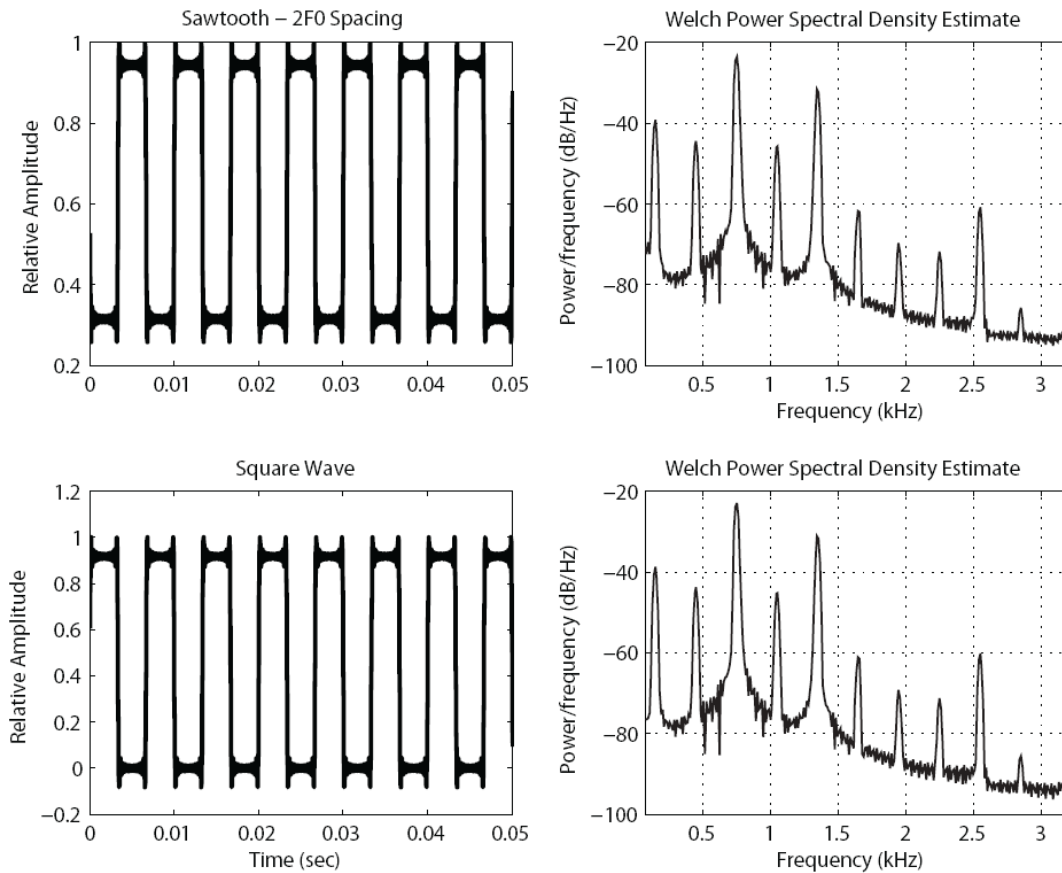


Figure 3. Visualization of timbral metamers for the /a/ (/a/) signals. Pictured are the source waveforms in the temporal domain as well as the filtered and RMS-equalized signals in the frequency domain. Sawtooth wave with every second harmonic removed (2F0 spacing) is on the top row; square wave on bottom. Source waveforms in the temporal domain are pictured prior to RMS equalization. Examination of the sawtooth wave with every second harmonic removed and the square wave signals demonstrates that though their frequency content is identical, there is a slight phase difference between the signals. Similarity of the signals in the frequency domain is clearly evident when examining the PSD estimate via Welch’s method. The peaks in the spectrum are identical between signals and since they are completely steady-state,

then their timbre is identical. We hypothesize that perception and cortical processing of the signals should likewise be identical.

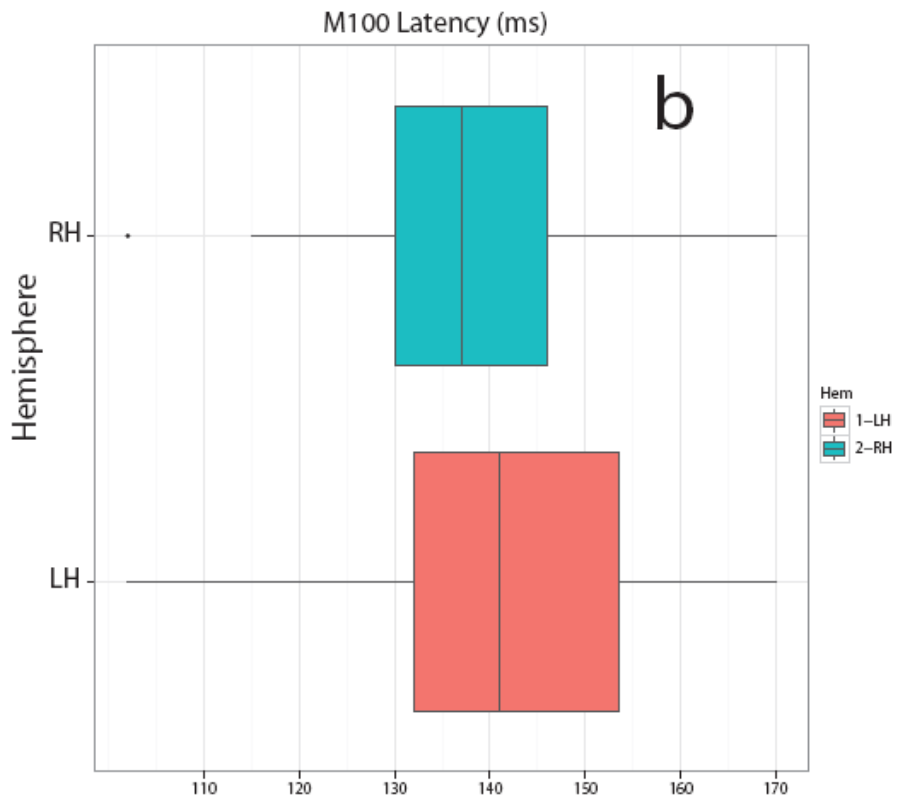
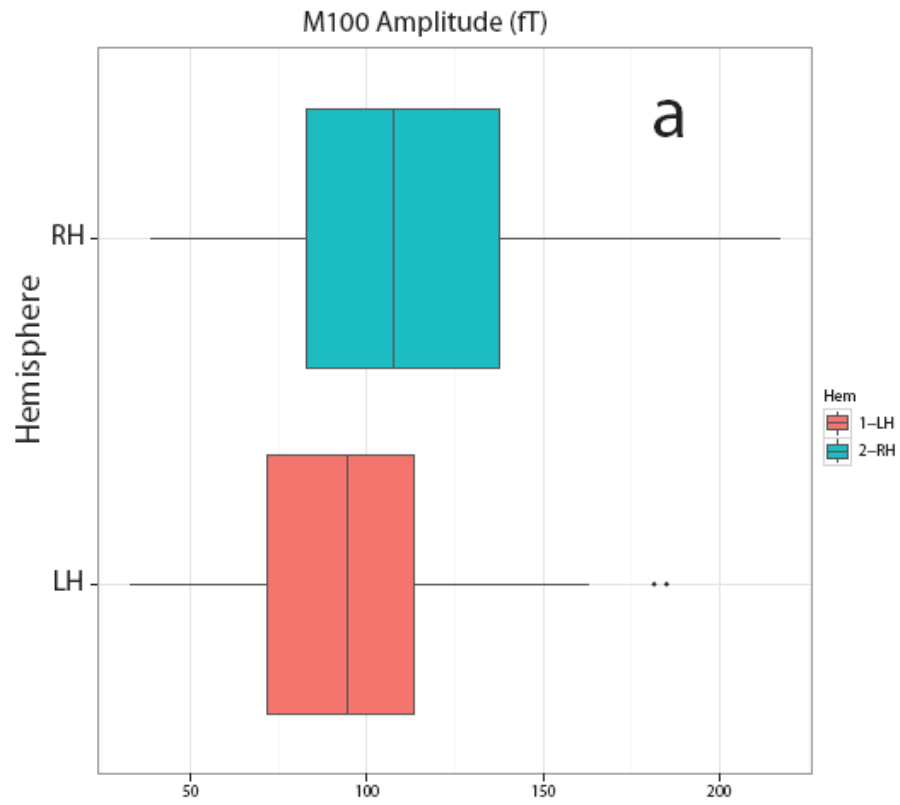


Figure 4. Boxplots illustrating normal approximations to pretest signals for M100 amplitude and latency, separated by hemisphere. RH is represented by the cyan box; LH magenta. Figure 4a illustrates the approximations for M100 amplitude, 4b for M100 latency. Though not apparent from the boxplots, the RH exhibited, on average, a larger amplitude and faster latency than the LH. This visualization also demonstrates the large variance observed across participants as they passively attended to the signals.

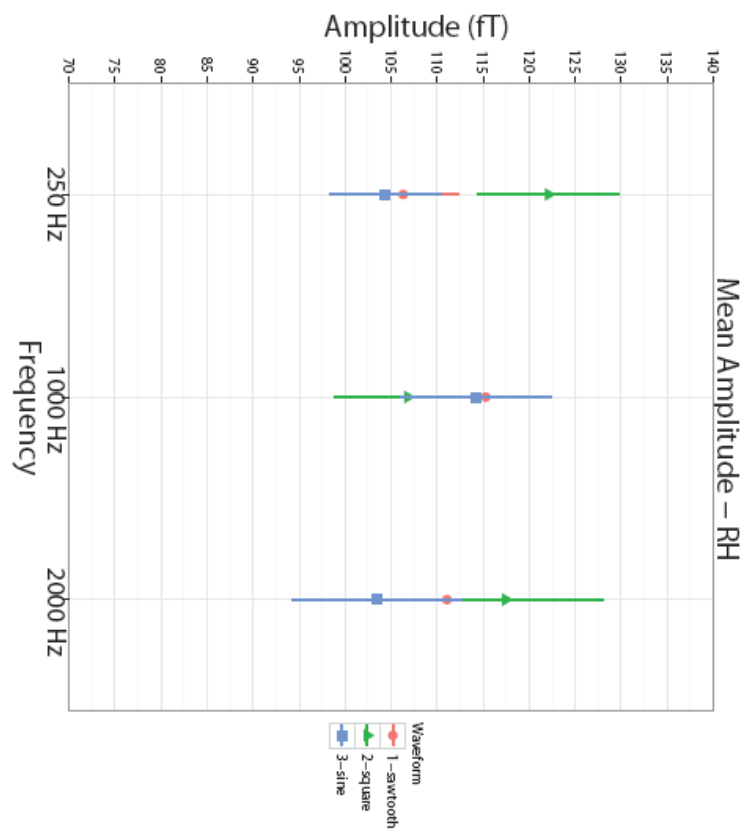
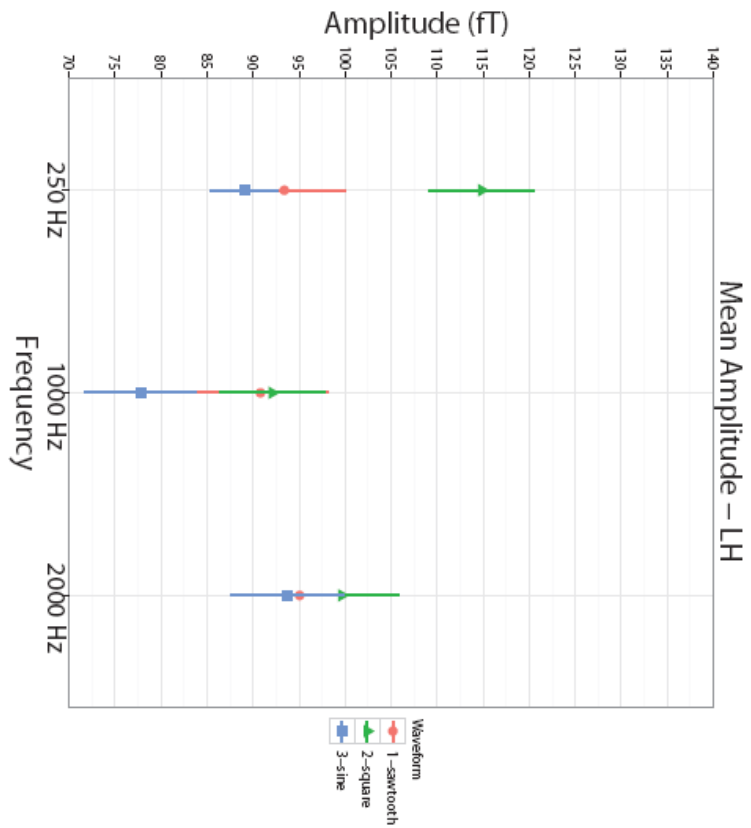


Figure 5. Mean M100 amplitude for each hemisphere, waveform and frequency for the experimental pretests. LH is on the left panel, RH on the right. Sawtooth waveforms are in magenta, square waveforms in green, sinusoidal waveforms in blue. Frequency increases from 250 Hz to 2000 Hz from L to R. From this visualization it is clear that the RH displays a greater M100 amplitude on average than the LH. Aside from a few isolated frequencies, no significant statistical difference between the waveforms at each frequency or overall was observed.

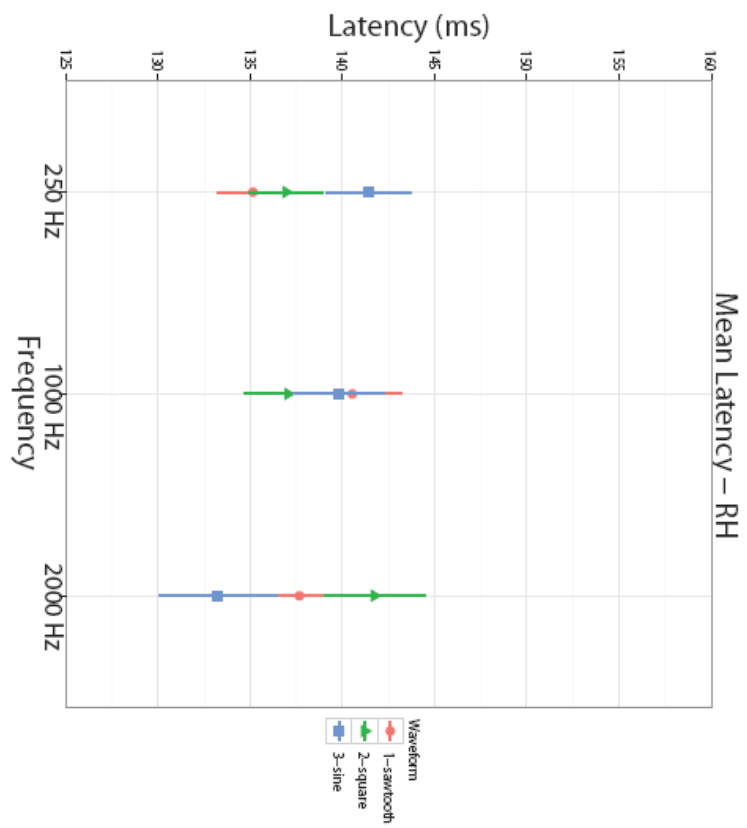
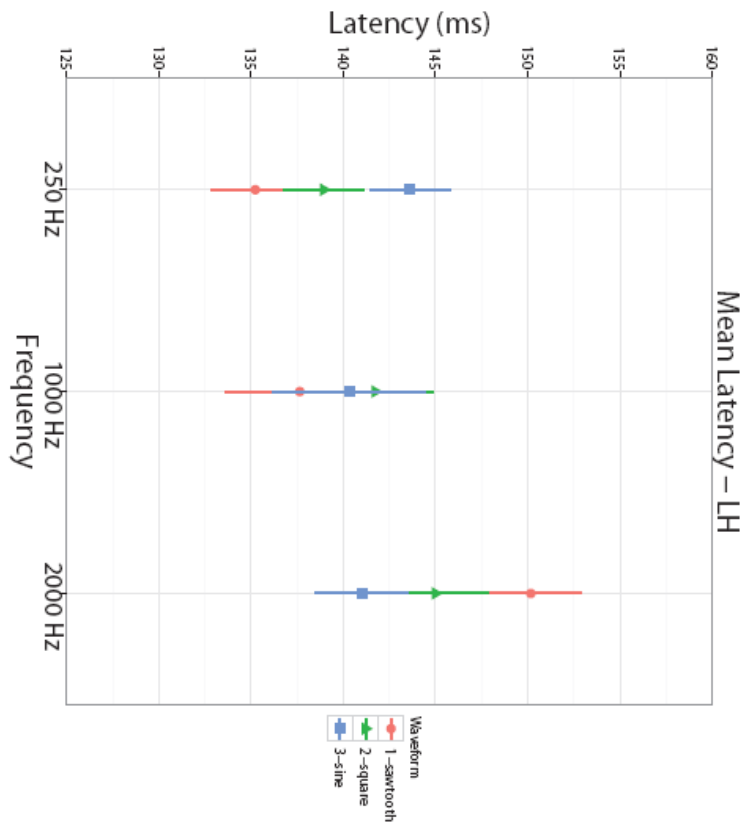


Figure 6. Mean M100 latency for each hemisphere, waveform and frequency for the experimental pretests. Conventions used are the same as in Figure 5. On average, the RH exhibited a faster latency than the LH. As with the M100 amplitude, aside from a few isolated cases, there was no consistent separation between each waveform and frequency.

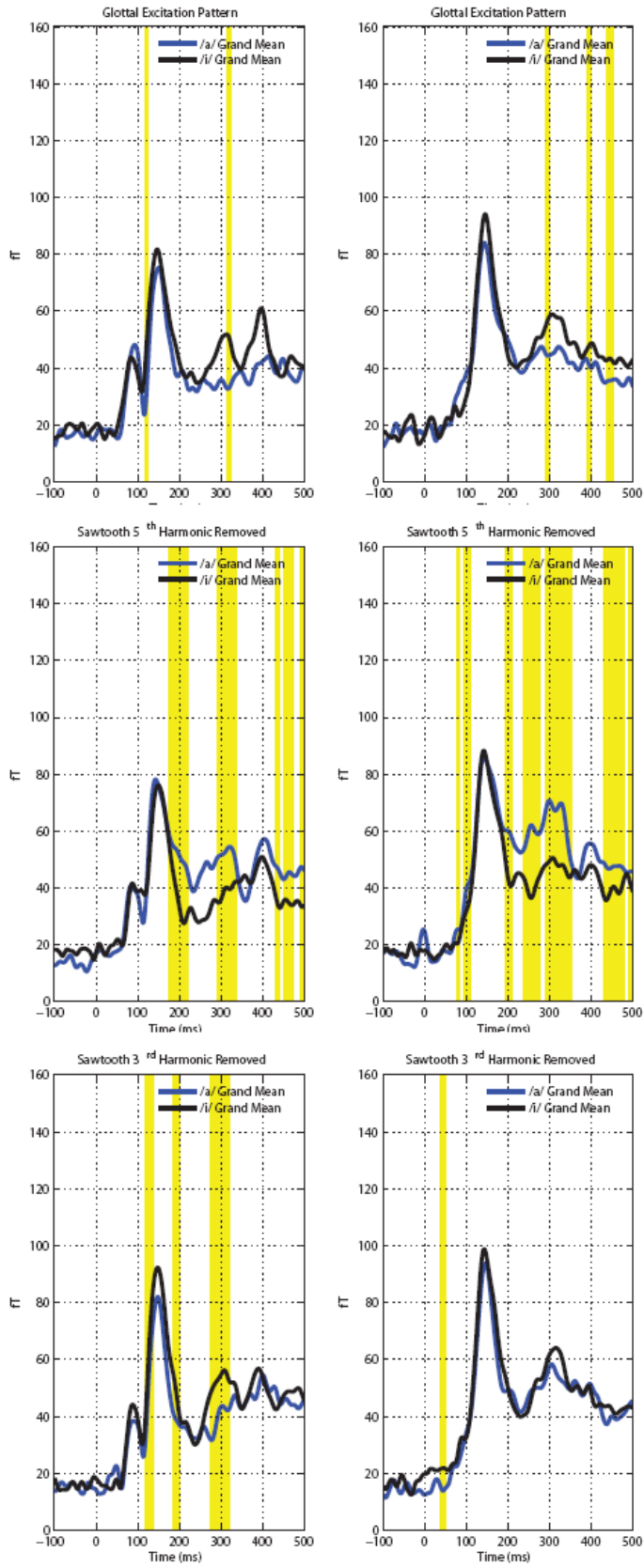


Figure 7. Grand averaged (RMS of RMS) waveforms for several conditions for the /ɑ/ - /i/ experimental assignment. The data are taken from the following source waveforms: (i) sawtooth wave approximating the glottal excitation pattern (top row), (ii) sawtooth wave with every fifth harmonic removed (middle row) and (iii) sawtooth wave with every third harmonic removed (bottom row). Data from the LH evoked responses are in the left column, RH data in the right column. Consecutive time points found to be statistically significant for evoked response amplitude differences are highlighted in yellow. Grand averages for the /ɑ/ signals are in blue; grand averages for the /i/ signals are in black. From the data visualizations, it is apparent that there is an asymmetry in the temporal evolution of the response; the LH consistently exhibits a peak (~90 msec) where a M50 response would be expected. The asymmetry in the M100 peak is also apparent; on average, the RH exhibited a larger M100 amplitude than the LH. Examination of the time points found to be significant via bootstrap resampling revealed that no time range was consistent across conditions.

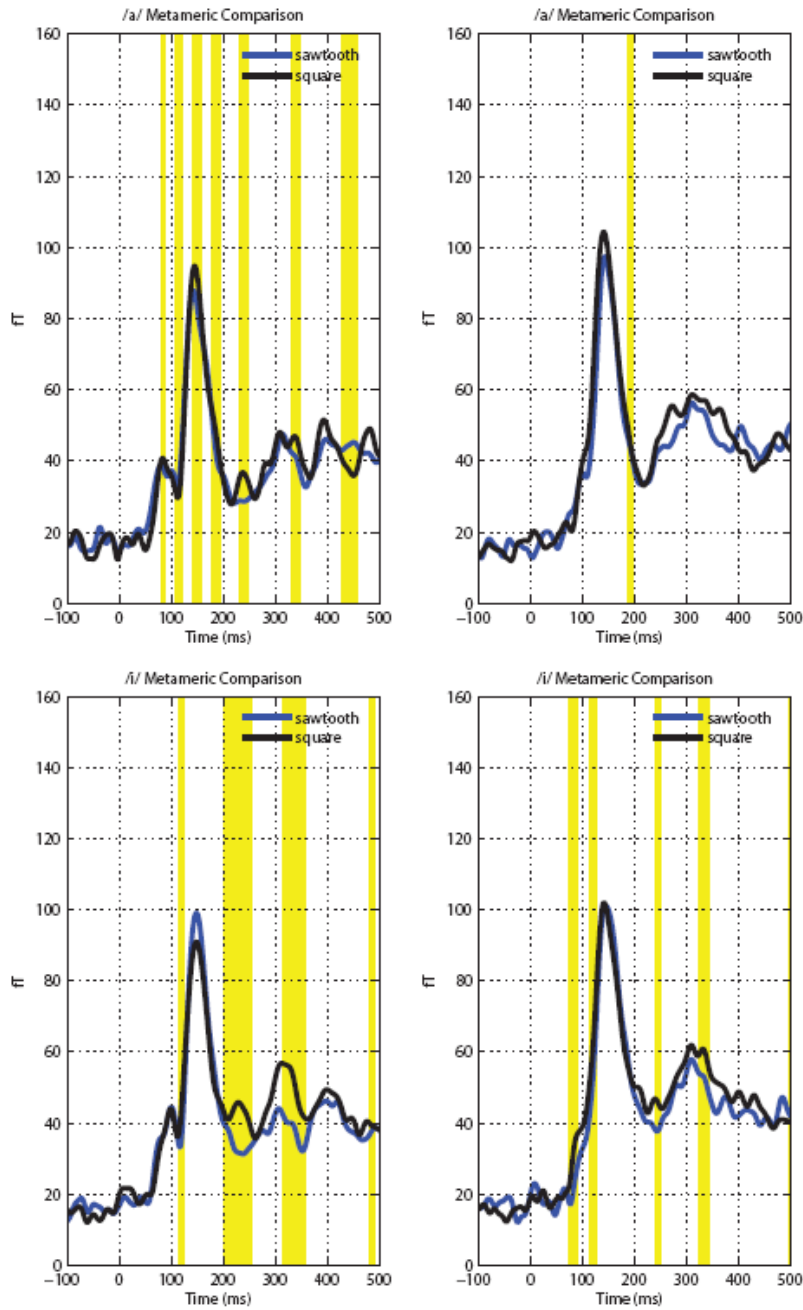


Figure 8. Grand averages (RMS of RMS) for the metameretic comparisons for each vowel, /a/ - /i/ experimental assignment. Consecutive time points found to be statistically significant for evoked response amplitude differences are highlighted in yellow. The sawtooth wave evoked response temporal evolution is plotted in blue;

square wave in black. In line with our hypotheses concerning the possible cortical processing of the metamer signals, the temporal evolution of the evoked responses of the metamers for each vowel are almost completely identical, especially in the time ranges for the M50 and M100 responses. However, since the differences in the evoked responses between the vowels was not significantly different nor were there any consistent differences regarding the time points found to be significant via bootstrap resampling, the functional implications of this observation are unclear.

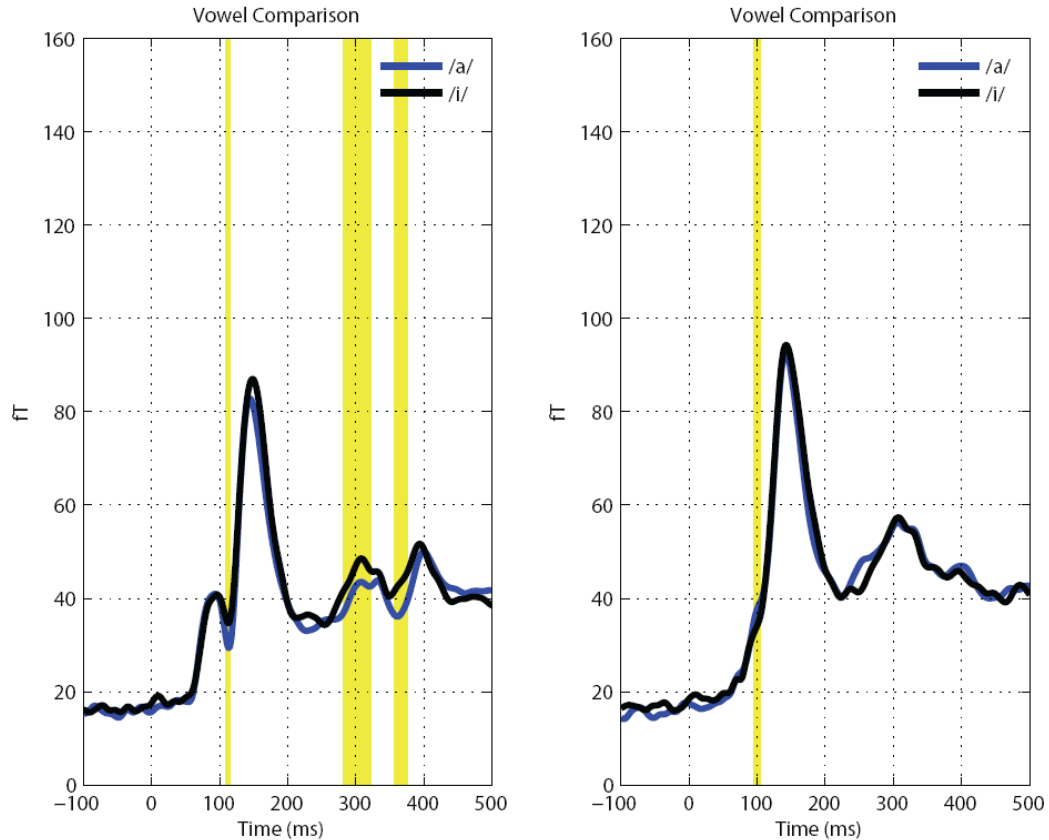
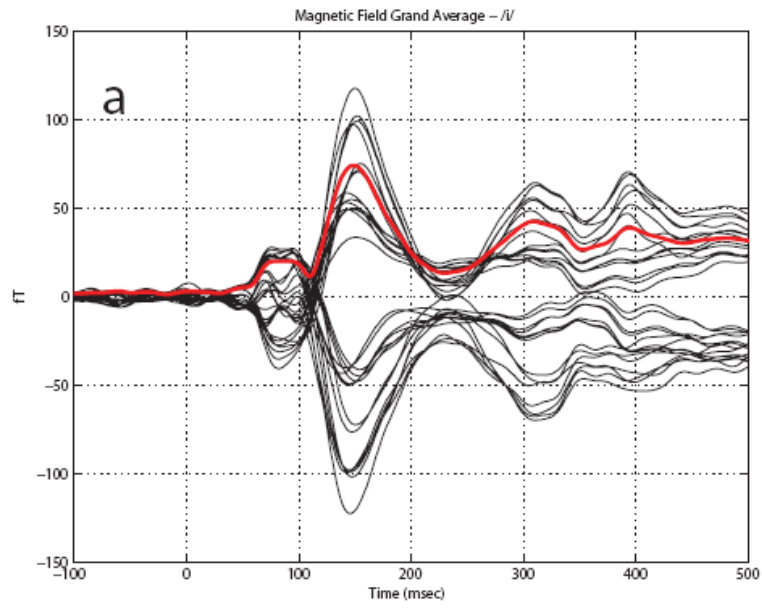
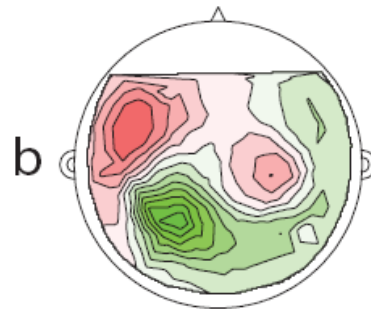


Figure 9. Grand averages (RMS of RMS) of the evoked responses to all vowel signals, collapsed across condition. Consecutive time points found to be statistically

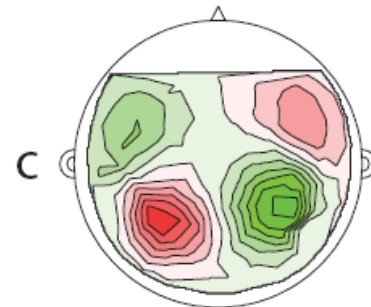
significant for evoked response amplitude differences are highlighted in yellow. /a/ evoked response grand averages are in blue; /i/ grand averages in black. In line with the statistical analyses of the mean evoked response values, there is very little difference between the /a/ and /i/ signal grand averages; in fact, the responses nearly overlap in the temporal domain. The asymmetry between the hemispheres regarding M50 generation and M100 amplitude and latency is clearly discerned; the LH M50 response is more robust. In spite of the near-identical responses, there are a series of time points near prior to M100 generation that are significant; though we did not perform dipole analyses, these time points may reflect differences in the source activity of the M100 generators. It should be noted the responses are somewhat late for a 'typical' M100; this may be a result of fundamental frequency. The fundamental frequency of the signals was 150 Hz; the M100 latencies observed (~140 msec) are typical of signals in that frequency range. The data suggest that M100 latency is thus a function spectral envelope and fundamental frequency.



M50 Topography



M100 Topography



Sustained Field Topography

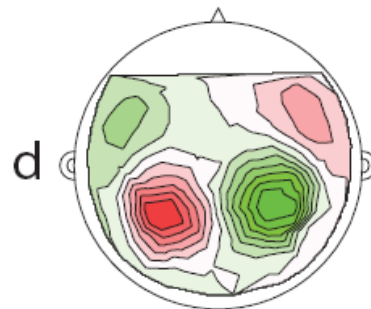


Figure 10. Grand average of the magnetic field deflections and RMS with grand averaged topography, /a/ - /i/ experimental assignment. Grand averages of the magnetic field deflection and resulting RMS in the temporal domain are visualized in Figure 10a; magnetic field topographies in 10c-d. Magnetic field deflections are plotted in black, RMS in red. For the magnetic field topographies, magnetic source is in red; magnetic sink in green. Data for the field deflections are taken from the grand averages of the /i/ signals across conditions from 28 channels common to all participants in both hemispheres. Channels/sensors selected were taken from individual subject data and were those found to record the largest field deflections as determined by the second experimental pretest. From this visualization of the data, it is clear that there are three distinct components of the evoked response: the M50 (~90 msec), M100 (~150 msec) and a sustained field component (~308 msec). The M50 response is much weaker than the M100 and exhibits a topography (sink-source distribution) opposite that of the M100 and sustained field components; it is also much more left-lateralized.

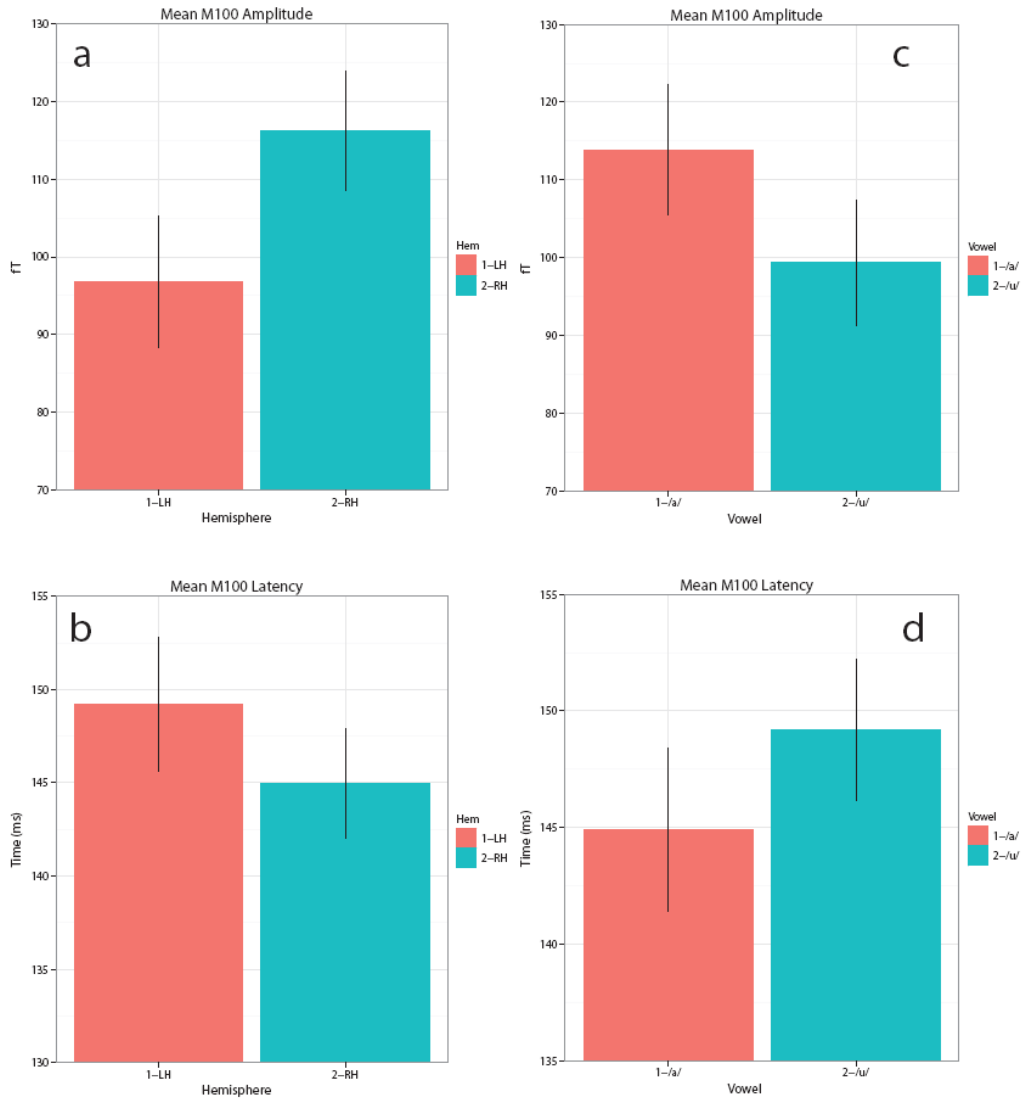


Figure 11. Bar plots of the mean M100 amplitude and latency and standard errors for the /a/ - /u/ experimental assignment. The data are separated as to hemisphere (11a,b) and vowel identity (11c,d). Hemispheric data is in the left column; vowel data in the right column. Mean M100 amplitude is on the top row; mean M100 latency on the bottom. Magenta bars indicate the LH and the /a/ signals; cyan the RH and /u/

signals. The data replicate previous findings, namely, that the evoked responses to the /u/ signals are slower and generate a smaller M100 amplitude.

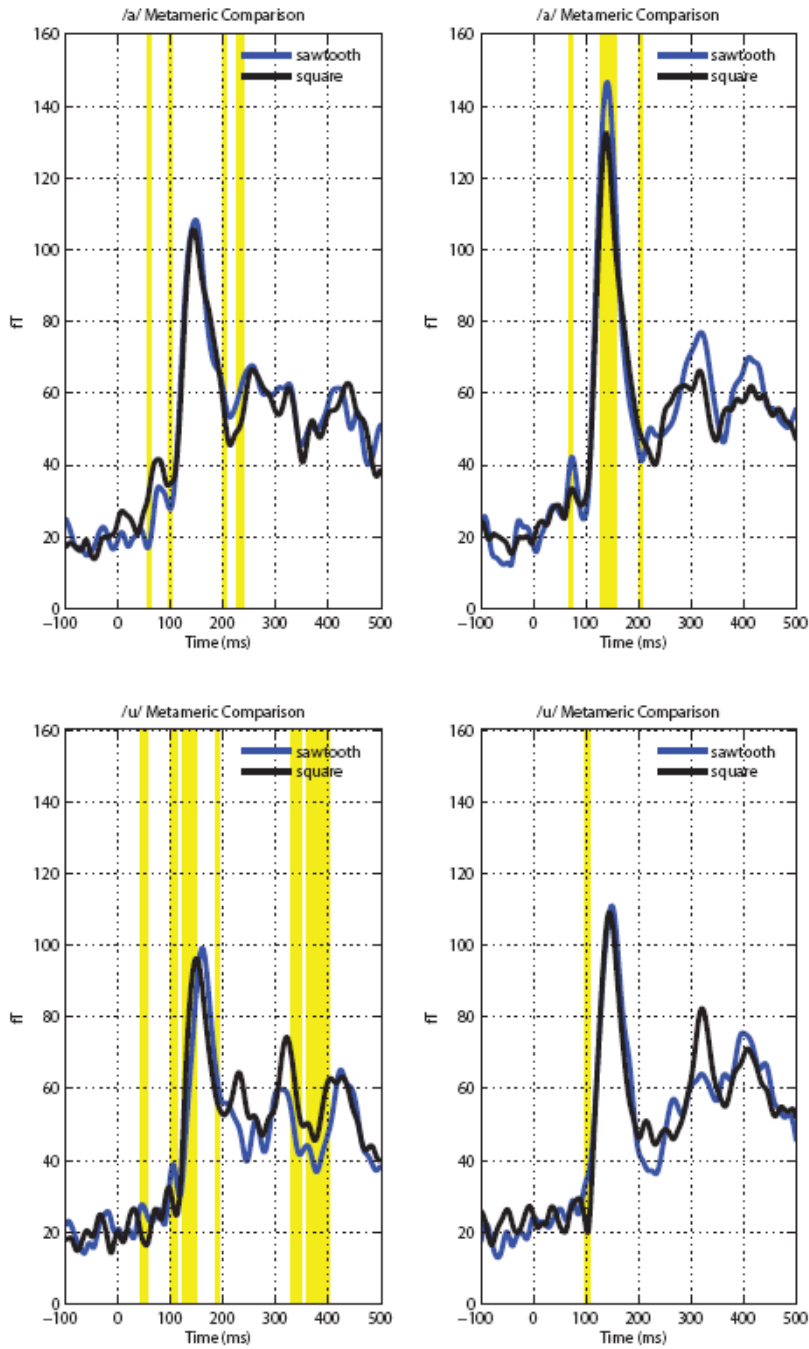


Figure 12. Grand averages (RMS of RMS) of the metamer comparisons for the /a/ - /u/ experimental assignment. Consecutive time points found to be statistically significant for evoked response amplitude differences are highlighted in yellow. The sawtooth wave evoked response temporal evolution is plotted in blue; square wave in black. As in the /a/ - /i/ assignment, the temporal evolution of the evoked responses for the metamers closely resemble one another. However, since the time points found to be significant for the difference in amplitudes were not consistent across conditions, it is not clear what the functional significance of the differences are.

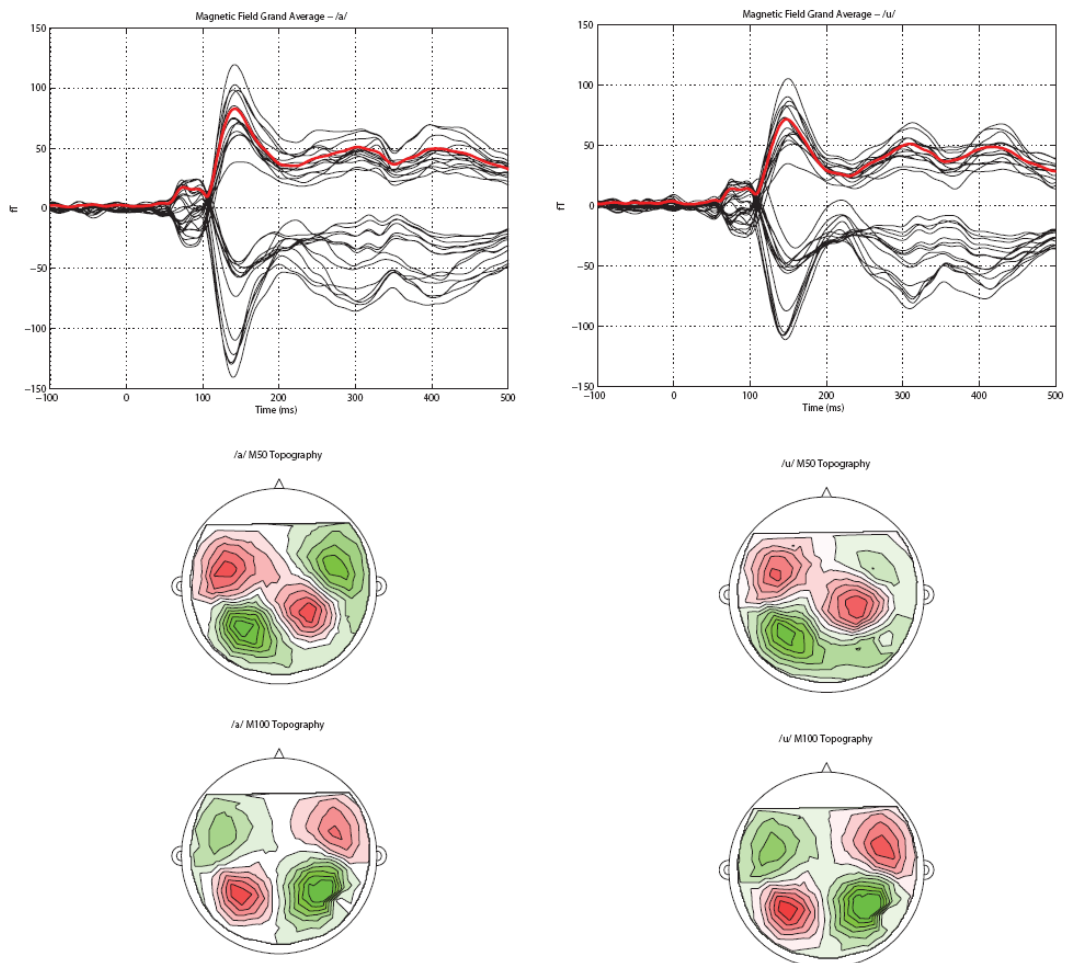


Figure 13. Grand average of the magnetic field deflections and RMS with grand averaged topography, /a/ - /u/ experimental assignment. Waveform and source-sink conventions are the same as in Figure 10. Grand averages to the /a/ signals are in the left column; grand averages for /u/ in the right column. The visualization illustrates the evoked responses to each vowel across conditions, along with the topographies for the M50 and M100 responses. The asymmetry between the hemispheres present in the /a/ - /i/ assignment is clearly visible here; though much weaker than the M100 the M50 exhibits a stronger response in the LH. The topographical difference between each component is also clearly delineated; the topographies are reversed between the M50 and M100 and sustained field, reflecting different patterns of neuronal activity.

Summary

The experiments presented cover three broad areas: physiological measurement of preattentive auditory processing, processing of audiovisual integration and signal discrimination. The first, third and fourth chapters share the most conceptually, while the second (audiovisual integration) provides data concerning the addition of visual information to an auditory signal. Together, all four chapters provide an account of the neuronal processing of signals that share important aspects of ecological signals in addition to the cognition of such signals.

The first chapter examines the processing of two-frequency tone complexes. The complexes consisted of a high frequency and a low frequency component, and the low frequency component was presented at one of five different loudness values. Both the tone complexes and their single frequency components were presented to the observers and the evoked response was evaluated using MEG. The M100 evoked component was consistently elicited and we observed that the complex tones exhibited a larger amplitude and faster latency than the single frequency tones at certain signal levels. The finding regarding the larger amplitude for the M100 was not unexpected; the complex signals contained more frequencies and a larger amount of energy than the single frequency signals. The larger amplitude might then be a consequence of driving a larger neuronal population. The finding concerning the faster M100 latency was however, unexpected, though not completely (Diesch and Luce, 1997; Lakatos et al., 2005). The data demonstrating the faster M100 latency is

a replication of previous data (Diesch and Luce, 1997), but several points confound the results. The first concerns to what extent the higher frequency component was driving the response. Frequencies in the 1-4 kHz range have lower detection thresholds than those in the ~100-500 Hz range, and this increased sensitivity may have been reflected in the M100 latency (Moore, 2004). A second possibility is that the amount of energy contained in the M100 analysis window (Gage and Roberts, 2000) may have resulted in the M100 latency facilitation (see, e.g., Sams et al., 1993). The increased energy in the M100 evaluation window for the complex signals may be analogous to that seen in pulse train data.

The last point, which relates the most to evaluation of the signals in the fourth chapter, is that the two-frequency complexes may significantly resemble ecological signals in some way, resulting in a faster and more robust response (Jacobsen et al., 2004). For example, vowels are mostly characterized by the spectral envelope structure, specifically the peaks (formants) in the spectral envelope (Plomp et al., 1967; Slawson, 1968; Klein et al., 1970; Hillenbrand et al., 1995; Diesch et al., 1996; Vihla and Salmelin, 2003; Jacobsen et al., 2004; Tiitinen et al., 2004; Mesgarani et al., 2008). The ratio of the first two formants especially, has been shown to be crucial to vowel identification and processing. The complex tones in the experiment share an obvious relationship to the formant structure of vowels in there are two distinct spectral peaks in the signal and these peaks may be evaluated in a manner similar to that of vowels. Additional analyses and experiments could analyze two-frequency complexes with spectral envelope structure similar to that of vowels (and within

critical bands) and source waveforms like those employed in the discrimination and MEG timbre experiments. The results of an experiment with different spectral peak distances and frequency values would provide data concerning frequency sensitivity and spectral peak evaluation. Another possible method of analyzing the signals is to calculate SCG as in other studies (Caclin et al., 2005) and see if that value is a predictor of evoked M100 amplitude and latency, when harmonic structural differences are present in a broad class of signals .

The third chapter examines the ability of an observer to discriminate between signals based on their spectral structure and the creation of hypothetical signal taxonomies and timbre spaces. The signals employed were synthesized in accordance with source-filter theory and were based on vowel token data (Hillenbrand et al., 1995; Fant et al., 2000). The filters corresponded to the transfer functions for the American English vowels /a/, /i/ and /u/. Source waveforms were based off of a sawtooth wave approximating the glottal excitation pattern and timbral differences were created by selectively removing harmonics. The set of source waveforms were as follows: a sawtooth waveform comprised of twenty-three harmonics, a sawtooth waveform with every fourth harmonic removed, a sawtooth waveform with every second harmonic removed, and a square wave where the maximum harmonic is the twenty-third. The sawtooth wave with every second harmonic removed and the square wave were the timbral metamers: signals that are physically different yet perceptually identical.

For the discrimination task, we hypothesized the metamers would be confused most often and that differences between non-metameric signals would be easily discriminated. The psychophysical data supported the metamer hypothesis; the metamers were confused the majority of the time. There was one surprising result concerning the proportion correct of the signals when analyzed by vowel category: the /a/ signals exhibited a much lower proportion correct than either the /i/ or /u/ signals. This may have been due to the transfer function of the /a/ signals. Specifically, the spectral peaks of the first two formants may have been ‘averaged’ in a sense; this would result in one overall peak as opposed to two distinct peaks and a ‘smearing’ of spectral information.

Possible signal taxonomies used a larger set of signals than in the discrimination task; the transfer functions were identical, but the source waveforms consisted of a sawtooth wave approximating the glottal excitation pattern, sawtooth waveforms with every sixth, fifth, fourth, third and second harmonic removed and a square wave where the maximum harmonic was the twenty-third. Possible taxonomies were created using hierarchical agglomerative clustering using the physical structure and PSD of the signals. Creation of possible timbre spaces was accomplished by transforming the signals into cochlear textures, applying spatial filters to highlight either gross or fine structure features and then applying MDS. We hypothesized the signals would be clustered according to vowel category (based on transfer function structure) and the metamer signals would be positioned closest to one another.

The results of the simple computational analyses demonstrated that it is possible to cluster signals according to mathematical properties describing them (Fellowes et al., 1997; Lakatos, 2000). Hierarchical clustering of the signals was largely based on a combination of the transfer function structure and source waveforms (for the physical structure) and transfer function structure for the PSD. For the MDS timbre space, the metamers were positioned closest to one another (this relationship was not entirely consistent for hierarchical clustering) and each of the coordinates and/or planes indexed either vowel category or some feature of spectral structure. Timbre space densities also differed according to transfer function; the /a/ signals were much less tightly clustered than the /i/ or /u/ signals. The results of these analyses are very promising and could be expanded in several ways. The first is to employ a greater variety of spatial filters for the texture analyses and increase the resolving power of the resulting texture. A second way is to employ more sophisticated methods; e.g., using a support vector machine as a pattern recognizer. The last way is to perhaps solve the one dimensional wave equation and cluster the resulting coefficients/representation. This would effectively constrain the boundary conditions for the problem and perhaps result in more informative timbre spaces.

The fourth chapter employs the identical signals to those used in the computational analyses to analyze preattentive timbre processing using MEG. Here, the hypotheses centered on how the latency and amplitude of the M50, M100 and P2m evoked components were modulated by the spectral structure of the signals. We

hypothesized the M50 would be elicited to their broadband nature, the M100 would exhibit sensitivity to spectral complexity, fundamental frequency and vowel transfer function structure, while the P2m would exhibit a larger amplitude to more spectrally complex signals. The most interesting analyses would be within vowel categories; we were curious as to how harmonic structure would be reflected in the evoked components. Specifically, we hypothesized the M100 latency (within vowel categories) would be faster for the more spectrally complex signals.

The data from the MEG timbre experiment indicate that the evoked response is not sufficient to characterize differences in harmonic structure. Rather, the M100 was the only component consistently elicited, and seemed to be sensitive to fundamental frequency and vowel transfer function (Diesch et al., 1996; Diesch and Luce, 1997; Obleser et al., 2003b; Obleser et al., 2003a, 2004; Tiitinen et al., 2004, 2005; Obleser et al., 2006). In other words, once in a specific latency regime for fundamental frequency, the M100 latency and amplitude was distinguished by vowel transfer function. However, even this processing had its limits: the latency and amplitude for /a/ and /i/ signals were statistically identical. Dipole analyses may reveal that the sources for the neuronal populations generating these responses are different, with the neurons generating the response to /i/ being located more anteriorly than the /a/ signals.

The paradigm in the MEG timbre experiment could be expanded by removing more harmonics from the source spectrum and analyzing to what extent spectral

envelope structure, specifically the ratio between the first two formants, affects M100 latency and amplitude. The data analysis could be improved by analyzing the induced response, using a time-frequency transform, since the evoked response does not seem to capture subtle differences in harmonic structure. Spectral center of gravity (Caclin et al., 2005) might also be a useful index frequency contributions to evoked response latency and amplitude.

The audiovisual experiment presented in the second chapter would seem to only bear passing resemblance to the other auditory-centric experiments. It fits in the larger narrative in the sense that it examines signals of an intermediate nature that capture crucial aspects of ecological signals. However, in the ecological speech signal, it is not typical to only process auditory information; visual information greatly facilitates signal processing (Sumbly and Pollack, 1954). The audiovisual experiment can then be viewed as exploring how redundancy across modalities facilitates processing (Molholm et al., 2002; Molholm et al., 2004).

The signals employed in the audiovisual experiment approximate the auditory and visual components of ecological speech. The visual component consists of an ellipse modulated as to resemble mouth movements while the auditory component consists of AM pink noise. The modulation rate (3.125 Hz) is consistent with those found in ecological speech signals (typically 2-16 Hz). The experiment uses the SSR and MEG as paradigms to explore how AV signals are processed in the spectral domain (i.e., privileged time scales) and how each component contributes to the

overall neuronal response. We hypothesized that the presence of redundancy resulted in a stronger SSR (as opposed to unimodal stimulation) and the difference in response power observed would be detected in the sensors overlying the parietal lobe.

The data demonstrated that an AV response was larger than a unimodal response in either modality and that the changes in the neuronal response did most likely arise in the parietal lobe. Our findings agree well with previous data concerning the evaluation of SSR and we have demonstrated for the first time entrainment to a unique set of bimodal signals (Chandrasekaran et al., 2009; Luo et al., 2010). The paradigm could be expanded by using ecological tokens as well as exploring more variable interactions by not using completely steady-state signals.

In sum, the experiments presented contribute to the literature concerning auditory and audiovisual processing by expanding on several experimental paradigms and using signals that of an intermediate nature to inform fundamental issues in cognition. The experiments not only replicate several previous findings, but also expand the literature by using novel signal sets in thereof the four experiments. Together, along with their possible iterations and extensions can aid in the understanding of auditory sensory processing, behavior and cognition, in regards to the processing of environmental and ecological signals.

Bibliography

- Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, Naumer MJ (2005) Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res* 166: 559-571.
- Auzou P, Eustache F, Etevenon P, Platel H, Rioux P, Lambert J, Lechevalier B, Zarifian E, Baron J (1995) Topographic EEG activations during timbre and pitch discrimination tasks using musical sounds. *Neuropsychologia* 33:25-37.
- Baayen R (2010) languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics" In, R package version 1.0 Edition.
- Baayen RH (2008) languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 0.953.
- Balzano GJ (1986) What Are Musical Pitch and Timbre? *Music Perception* 3:297-314.
- Baumann O, Greenlee MW (2007) Neural Correlates of Coherent Audiovisual Motion Perception. *Cereb Cortex* 17: 1433-1443.
- Belin P, Fecteau S, Bédard C (2004) Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 8:129-135.
- Belin P, Zatorre R (2000) 'What', 'where' and 'how' in auditory cortex. *Nat Neurosci* 3:965-966.
- Belin P, Zatorre R (2003) Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14:2105-2109.

- Belin P, Zatorre R, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403:309-312.
- Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur J Neurosci* 20: 2225-2234.
- Bigand E, Pineau M (1997) Global context effects on musical expectancy. *Percept Psychophys* 59:1098-1107.
- Binder J, Liebenthal E, Possing E, Medler D, Ward B (2004) Neural correlates of sensory and decision processes in auditory object identification. *Nat Neurosci* 7:295-301.
- Boucher R, Bryden MP (1997) Laterality effects in the processing of melody and timbre. *Neuropsychologia* 35:1467-1473.
- Brattico E, Tervaniemi M, Picton T (2003) Effects of brief discrimination-training on the auditory N1 wave. *Neuroreport* 14:2489-2492.
- Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts: The MIT Press.
- Bregman AS, Pinker S (1978) Auditory streaming and the building of timbre. *Can J Psychol* 32:19-31.
- Bruckert L, Bestelmeyer P, Latinus M, Rouger J, Charest I, Rousselet GA, Kawahara H, Belin P (2010) Vocal attractiveness increases by averaging. *Curr Biol* 20:116-120.
- Caclin A, Brattico E, Tervaniemi M, Näätänen R, Morlet D, Giard M, McAdams S (2006) Separate neural processing of timbre dimensions in auditory sensory memory. *J Cogn Neurosci* 18:1959-1972.

- Caclin A, Giard M, Smith B, McAdams S (2007) Interactive processing of timbre dimensions: a Garner interference study. *Brain Res* 1138:159-170.
- Caclin A, McAdams S, Smith B, Giard M (2008) Interactive processing of timbre dimensions: an exploration with event-related potentials. *J Cogn Neurosci* 20:49-64.
- Caclin A, McAdams S, Smith B, Winsberg S (2005) Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. *J Acoust Soc Am* 118:471-482.
- Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999) Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10: 2619-2623.
- Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology* 10: 649-657.
- Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001) Detection of Audiovisual Integration Sites in Humans by Application of Electrophysiological Criteria to the BOLD effect. *NeuroImage* 14: 427-438.
- Campbell R (2008) The processing of audiovisual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci* 363: 1001-1010.
- Canty A, Ripley B (2010) boot: Bootstrap R (S-Plus) Functions. R package version 1.2-42

- Cappe C, Thut G, Romei V, Murray MM (2010) Auditory-Visual Multisensory Interactions in Humans: Timing, Topography, Directionality, and Sources. *J Neurosci* 30: 12572-12580.
- Carhart R, Jerger JF (1959). Preferred method for clinical determination of pure-tone thresholds. *J Speech Hear Dis*, 24:330-345.
- Chait M, de Cheveigné A, Poeppel D, Simon JZ (2010) Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia* 48:3262-3271.
- Chait M, Poeppel D, de Cheveigné A, Simon J (2007) Processing asymmetry of transitions between order and disorder in human auditory cortex. *J Neurosci* 27:5207-5214.
- Chait M, Poeppel D, Simon J (2006) Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cereb Cortex* 16:835-848.
- Chait M, Simon J, Poeppel D (2004) Auditory M50 and M100 responses to broadband noise: functional implications. *Neuroreport* 15:2455-2458.
- Chandrasekaran C, Trubanova A, Stillitano S, Caplier A, Ghazanfar AA (2009) The Natural Statistics of Audiovisual Speech. *PLoS Comput Biol* 5: e1000436.
- Chartrand J, Belin P (2006) Superior voice timbre processing in musicians. *Neurosci Lett* 405:164-167.
- Chi T, Shamma S NSL Toolbox. <http://www.isr.umd.edu/Labs/NSL/Software.htm>
- Clarkson M, Clifton R (1995) Infants' pitch perception: inharmonic tonal complexes. *J Acoust Soc Am* 98:1372-1379.

- Cleveland T (1977) Acoustic properties of voice timbre types and their influence on voice classification. *J Acoust Soc Am* 61:1622-1629.
- Cristia A, Seidl A (2008) Is Infants' Learning of Sound Patterns Constrained by Phonological Features? *Language Learning and Development* 4:203-227.
- Crummer G, Walton J, Wayman J, Hantz E, Frisina R (1994) Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *J Acoust Soc Am* 95:2720-2727.
- Culling J, Darwin C (1993) The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Percept Psychophys* 54:303-309.
- Darwin C, Gardner R (1986) Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. *J Acoust Soc Am* 79:838-845.
- Darwin C, Pattison H, Gardner R (1989) Vowel quality changes produced by surrounding tone sequences. *Percept Psychophys* 45:333-342.
- Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge ; New York, NY, USA: Cambridge University Press.
- de Cheveigné A, Simon JZ (2007) Denoising based on time-shift PCA. *J Neurosci Methods* 165: 297-305.
- Deutsch D (1999) *The psychology of music*, 2nd Edition. San Diego: Academic Press.
- Diehl RL (1981) Feature detectors for speech: a critical reappraisal. *Psychol Bull* 89:1-18.
- Diehl RL (2008) Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Philos Trans R Soc Lond B Biol Sci* 363:965-978.

- Diesch E, Eulitz C, Hampson S, Ross B (1996) The neurotopography of vowels as mirrored by evoked magnetic field measurements. *Brain Lang* 53:143-168.
- Diesch E, Luce T (1997) Magnetic fields elicited by tones and vowel formants reveal tonotopy and nonlinear summation of cortical activation. *Psychophysiology* 34:501-510.
- Diesch E, Luce T (2000) Topographic and temporal indices of vowel spectral envelope extraction in the human auditory cortex. *J Cogn Neurosci* 12:878-893.
- Dietrich V, Nieschalk M, Stoll W, Rajan R, Pantev C (2001) Cortical reorganization in patients with high frequency cochlear hearing loss. *Hear Res* 158:95-101.
- Dissard P, Darwin C (2001) Formant-frequency matching between sounds with different bandwidths and on different fundamental frequencies. *J Acoust Soc Am* 110:409-415.
- Dobie RA, Wilson MJ (1996) A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion product otoacoustic emissions, or other sinusoids. *J Acoust Soc Am* 100: 2236-2246.
- Driver J, Spence C (1998) Crossmodal attention. *Curr Opin Neurobiol* 8: 245-253.
- Eggermont JJ (1995) Evoked potentials as indicators of auditory development. *Int J Pediatr Otorhinolaryngol* 32 Suppl:S183-186.
- Eulitz C, Diesch E, Pantev C, Hampson S, Elbert T (1995) Magnetic and electric brain activity evoked by the processing of tone and vowel stimuli. *J Neurosci* 15:2748-2755.

- Fant G (1972) Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR* 13:028-052.
- Fant G (1980) The relations between area functions and the acoustic signal. *Phonetica* 37:55-86.
- Fant G, Kruckenberg A, Liljencrants J (2000) The source-filter frame of prominence. *Phonetica* 57:113-127.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton.
- Fellowes J, Remez R, Rubin P (1997) Perceiving the sex and identity of a talker without natural vocal timbre. *Percept Psychophys* 59:839-849.
- Fisher NI (1996) *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.
- Forster KI, Forster JC (2003) DMDX: a windows display program with millisecond accuracy. *Behav Res Methods Instrum Comput* 35:116-124.
- Fort A, Delpuech C, Pernier J, Giard M-H (2002) Dynamics of Cortico-subcortical Crossmodal Operations Involved in Audiovisual Object Detection in Humans. *Cereb Cortex* 12: 1031-1039.
- Fujioka T, Ross B, Okamoto H, Takeshima Y, Kakigi R, Pantev C (2003) Tonotopic representation of missing fundamental complex sounds in the human auditory cortex. *Eur J Neurosci* 18:432-440.
- Gage N, Roberts T (2000) Temporal integration: reflections in the M100 of the auditory evoked field. *Neuroreport* 11:2723-2726.
- Gander PE, Bosnyak DJ, Roberts LE (2010) Evidence for modality-specific but not

- frequency specific modulation of human primary auditory cortex by attention. *Hear Res* 268: 213-226.
- Gfeller K, Knutson J, Woodworth G, Witt S, DeBus B (1998) Timbral recognition and appraisal by adult cochlear implant users and normal-hearing adults. *J Am Acad Audiol* 9:1-19.
- Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? *Trends Cogn Sci* 10: 278-285.
- Godey B, Schwartz D, de Graaf JB, Chauvel P, Liégeois-Chauvel C (2001) Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. *Clin Neurophysiol* 112:1850-1859.
- Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am* 108: 1197-1208.
- Grey JM (1977) Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America* 61:1270-1277.
- Grey JM, Gordon JW (1978) Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America* 63:1493-1500.
- Hall JW 3rd (1977) Elements of timbre perception. *TIT J Life Sci* 7:43-51.
- Hari R, Aittoniemi K, Järvinen M, Katila T, Varpula T (1980) Auditory evoked transient and sustained magnetic fields of the human brain. Localization of neural generators. *Exp Brain Res* 40:237-240.
- Hari R, Kaila K, Katila T, Tuomisto T, Varpula T (1982) Interstimulus interval dependence of the auditory vertex response and its magnetic counterpart:

- implications for their neural generation. *Electroencephalogr Clin Neurophysiol* 54:561-569.
- Hari R, Mäkelä J (1988) Modification of neuromagnetic responses of the human auditory cortex by masking sounds. *Exp Brain Res* 71:87-92.
- Hartmann WM (1998) *Signals, Sound, and Sensation*, Corrected fifth printing, 2005 Edition. New York: Springer Science+Business Media, LLC.
- Hennig C (2010) fpc: Flexible procedures for clustering. R package version 2.0-3.
- Hershenson M (1962) Reaction time as a measure of intersensory facilitation. *J Exp Psychol* 63: 289.
- Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends Cogn Sci* 4:131-138.
- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393-402.
- Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. *J Acoust Soc Am* 97:3099-3111.
- Hirata Y, Kuriki S, Pantev C (1999) Musicians with absolute pitch show distinct neural activities in the auditory cortex. *Neuroreport* 10:999-1002.
- Howard M, Poeppel D (2009) Hemispheric asymmetry in mid and long latency neuromagnetic responses to single clicks. *Hear Res* 257:41-52.
- Howard MF, Poeppel D (2010) Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol* 104:2500-2511.

- Inui K, Okamoto H, Miki K, Gunji A, Kakigi R (2006) Serial and parallel processing in the human auditory cortex: a magnetoencephalographic study. *Cereb Cortex* 16:18-30.
- Iverson P, Krumhansl C (1993) Isolating the dynamic attributes of musical timbre. *J Acoust Soc Am* 94:2595-2603.
- Jacobsen T, Schröger E, Alter K (2004a) Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology* 41:654-659.
- Jacobsen T, Schröger E, Sussman E (2004b) Pre-attentive categorization of vowel formant structure in complex tones. *Brain Res Cogn Brain Res* 20:473-479.
- Jenkins J 3rd, Idsardi WJ, Poeppel D (2010) The Analysis of Simple and Complex Auditory Signals in Human Auditory Cortex: Magnetoencephalographic Evidence From M100 Modulation. *Ear and Hearing* 31: 515-526.
- Jones EG, Powell TP (1970) An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain* 93: 793-820.
- Jones S, Perez N (2001) The auditory 'C-process': analyzing the spectral envelope of complex sounds. *Clin Neurophysiol* 112:965-975.
- Kayser C, Petkov CI, Logothetis N, K. (2008) Visual Modulation of Neurons in Auditory Cortex. *Cereb Cortex* 18: 1560-1574.
- Kelly SP, Gomez-Ramirez M, Foxe JJ (2008) Spatial Attention Modulates Initial Afferent Activity in Human Primary Visual Cortex. *Cereb Cortex* 18: 2629-2636.

- Kendall RA, Carterette EC (1993) Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives. *Music Perception: An Interdisciplinary Journal* 10:445-467.
- Kirveskari E, Salmelin R, Hari R (2006) Neuromagnetic responses to vowels vs. tones reveal hemispheric lateralization. *Clin Neurophysiol* 117:643-648.
- Kruskal J (1964a) Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* 29:1-27.
- Kruskal J (1964b) Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika* 29:115-129.
- Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. *Science* 320:110.
- Lakatos P, Pincze Z, Fu K, Javitt D, Karmos G, Schroeder C (2005) Timing of pure tone and noise-evoked responses in macaque auditory cortex. *Neuroreport* 16:933-937.
- Lakatos S (2000) A common perceptual space for harmonic and percussive timbres. *Percept Psychophys* 62:1426-1439.
- Lakatos S, McAdams S, Caussé R (1997) The representation of auditory source characteristics: simple geometric form. *Percept Psychophys* 59:1180-1190.
- Lalor EC, Kelly SP, Pearlmutter BA, Reilly RB, Foxe JJ (2007) Isolating endogenous visuo-spatial attentional effects using the novel visual-evoked spread spectrum analysis (VESPA) technique. *Eur J Neurosci* 26: 3536--3542.
- Lewicki M (2002) Efficient coding of natural sounds. *Nat Neurosci* 5:356-363.

- Loizou P, Mani A, Dorman M (2003) Dichotic speech recognition in noise using reduced spectral cues. *J Acoust Soc Am* 114:475-483.
- Luo H, Boemio A, Gordon M, Poeppel D (2007) The perception of FM sweeps by Chinese and English listeners. *Hear Res* 224:75-83.
- Luo H, Liu Z, Poeppel D (2010) Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using Low-Frequency Neuronal Phase Modulation. *PLoS Biol* 8: e1000445.
- Luo H, Poeppel D (2007) Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron* 54: 1001-1010.
- Luo H, Wang Y, Poeppel D, Simon JZ (2006) Concurrent Encoding of Frequency and Amplitude Modulation in Human Auditory Cortex: MEG Evidence. *J Neurophysiol* 96: 2712-2723.
- Lütkenhöner B (2003) Single-dipole analyses of the N100m are not suitable for characterizing the cortical representation of pitch. *Audiol Neurootol* 8:222-233.
- Lütkenhöner B, Klein J (2007) Auditory evoked field at threshold. *Hear Res* 228:188-200.
- Lütkenhöner B, Lammertmann C, Knecht S (2001) Latency of auditory evoked field deflection N100m ruled by pitch or spectrum? *Audiol Neurootol* 6:263-278.
- Lütkenhöner B, Seither-Preisler A, Seither S (2006) Piano tones evoke stronger magnetic fields than pure tones or noise, both in musicians and non-musicians. *Neuroimage* 30:927-937.

- Lütkenhöner B, Steinsträter O (1998) High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiol Neurootol* 3:191-213.
- Lütkenhöner, B. & Poeppel, D. (in press). From tones to speech. In Schreiner, C. & J. Winer (Eds.), *Auditory Cortex*. New York: Springer.
- Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci* 28: 264-271.
- Macmillan NA, Creelman CD (1991) *Detection theory : a user's guide*. Cambridge [England] ; New York: Cambridge University Press.
- Mäkelä A, Alku P, May P, Mäkinen V, Tiitinen H (2004) The auditory n100m response reflects changes in speech fundamental frequency. *Neurol Clin Neurophysiol* 2004:49.
- Mäkinen V, May P, Tiitinen H (2004) Transient brain responses predict the temporal dynamics of sound detection in humans. *Neuroimage* 21:701-706.
- Marozeau J, de Cheveigné A (2007) The effect of fundamental frequency on the brightness dimension of timbre. *J Acoust Soc Am* 121:383-387.
- Marozeau J, de Cheveigné A, McAdams S, Winsberg S (2003) The dependency of timbre on fundamental frequency. *J Acoust Soc Am* 114:2946-2957.
- MATLAB (2009) Version R2009a Natick, MA The Mathworks.
- MATLAB (2005) Version 7, SP14. Natick, MA: The Mathworks.
- McAdams S, Winsberg S, Donnadiou S, De Soete G, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol Res* 58:177-192.

- Melara R, Marks L (1990a) Perceptual primacy of dimensions: support for a model of dimensional interaction. *J Exp Psychol Hum Percept Perform* 16:398-414.
- Melara R, Marks L (1990b) Interaction among auditory dimensions: timbre, pitch, and loudness. *Percept Psychophys* 48:169-178.
- Mesgarani N, David S, Fritz J, Shamma S (2008) Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am* 123:899-909.
- Mesulam MM (1998) From sensation to cognition. *Brain* 121: 1013-1052.
- Meyer M, Baumann S, Jancke L (2006) Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *Neuroimage* 32:1510-1523.
- Miller BT, D'Esposito M (2005) Searching for "the Top" in Top-Down Control. *Neuron* 48: 535-538.
- Miller J (1989) Auditory-perceptual interpretation of the vowel. *J Acoust Soc Am* 85:2114-2134.
- Mizuochi T, Yumoto M, Karino S, Itoh K, Yamakawa K, Kaga K (2005) Perceptual categorization of sound spectral envelopes reflected in auditory-evoked N1m. *Neuroreport* 16:555-558.
- Mizuochi T, Yumoto M, Karino S, Itoh K, Yamasoba T (2007) Latency variation of auditory N1m responses to vocal and nonvocal sounds. *Neuroreport* 18:1945-1949.
- Molholm S, Martinez A, Shpaner M, Foxe JJ (2007) Object-based attention is multisensory: co-activation of an object's representations in ignored sensory modalities. *Eur J Neurosci* 26: 499-509.

- Molholm S, Ritter W, Javitt DC, Foxe JJ (2004) Multisensory Visual-Auditory Object Recognition in Humans: a High-density Electrical Mapping Study. *Cereb Cortex* 14: 452-465.
- Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research* 14: 115-128.
- Molholm S, Sehatpour P, Mehta AD, Shpaner M, Gomez-Ramirez M, Ortigue S, Dyke JP, Schwartz TH, Foxe JJ (2006) Audiovisual Multisensory Integration in Superior Parietal Lobule Revealed by Human Intracranial Recordings. *J Neurophysiol* 96: 721-729.
- Monahan P, de Souza K, Idsardi W (2008) Neuromagnetic evidence for early auditory restoration of fundamental pitch. *PLoS One* 3:e2900.
- Moore B (2003) Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otol Neurotol* 24:243-254.
- Moore BC (2004) *An Introduction to the Psychology of Hearing*, Fifth Edition. New York: Elsevier Ltd.
- Moore BCJ (1995) *Hearing*. San Diego: Academic Press.
- Moore, B. C. J. (2004). The Perception of Loudness. In *An Introduction to the Psychology of Hearing* (pp. 127–162). San Diego, CA: Elsevier, Academic Press.
- Müller MM, Teder W, Hillyard SA (1997) Magnetoencephalographic recording of steady-state visual evoked cortical activity. *Brain Topography* 9: 163-168.

- Murray M, Camen C, Gonzalez Andino S, Bovet P, Clarke S (2006) Rapid brain discrimination of sounds of objects. *J Neurosci* 26:1293-1302.
- Murray MM, Foxe JJ, Wylie GR (2005) The brain uses single-trial multisensory memories to discriminate without awareness. *NeuroImage* 27: 473-478.
- Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. *Curr Opin Neurobiol* 14:474-480.
- Nikjeh DA, Lister JJ, Frisch SA (2009) Preattentive cortical-evoked responses to pure tones, harmonic tones, and speech: influence of music training. *Ear Hear* 30:432-446.
- Obleser J, Boecker H, Drzezga A, Haslinger B, Hennenlotter A, Roettinger M, Eulitz C, Rauschecker J (2006) Vowel sound extraction in anterior superior temporal cortex. *Hum Brain Mapp* 27:562-571.
- Obleser J, Eisner F, Kotz S (2008) Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *J Neurosci* 28:8116-8123.
- Obleser J, Elbert T, Lahiri A, Eulitz C (2003b) Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Brain Res Cogn Brain Res* 15:207-213.
- Obleser J, Kotz S (2010) Expectancy constraints in degraded speech modulate the language comprehension network. *Cereb Cortex* 20:633-640.
- Obleser J, Lahiri A, Eulitz C (2003a) Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset. *Neuroimage* 20:1839-1847.

- Obleser J, Lahiri A, Eulitz C (2004a) Magnetic brain response mirrors extraction of phonological features from spoken vowels. *J Cogn Neurosci* 16:31-39.
- Obleser J, Rockstroh B, Eulitz C (2004b) Gender differences in hemispheric asymmetry of syllable processing: left-lateralized magnetic N100 varies with syllable categorization in females. *Psychophysiology* 41:783-788.
- Obleser J, Wise R, Alex Dresner M, Scott S (2007) Functional integration across brain regions improves speech perception under adverse listening conditions. *J Neurosci* 27:2283-2289.
- Okamoto H, Stracke H, Bermudez P, Pantev C (2010) Sound Processing Hierarchy within Human Auditory Cortex. *J Cogn Neurosci*.
- Okamoto H, Stracke H, Pantev C (2008) Neural interactions within and beyond the critical band elicited by two simultaneously presented narrow band noises: a magnetoencephalographic study. *Neuroscience* 151:913-920.
- Okamoto H, Stracke H, Zwitserlood P, Roberts LE, Pantev C (2009) Frequency-specific modulation of population-level frequency tuning in human auditory cortex. *BMC Neurosci* 10:1.
- Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9: 97-113.
- Olson IR, Gatenby JC, Gore JC (2002) A comparison of bound and unbound audiovisual information processing in the human cerebral cortex. *Cognitive Brain Research* 14: 129-138.

- Overath T, Kumar S, Stewart L, von Kriegstein K, Cusack R, Rees A, Griffiths TD (2010) Cortical mechanisms for the segregation and representation of acoustic textures. *J Neurosci* 30:2070-2076.
- Pantev C, Bertrand O, Eulitz C, Verkindt C, Hampson S, Schuierer G, Elbert T (1995) Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalogr Clin Neurophysiol* 94:26-40.
- Pantev C, Hoke M, Lütkenhöner B, Lehnertz K (1989) Tonotopic organization of the auditory cortex: pitch versus frequency representation. *Science* 246:486-488.
- Pardo JS, Fowler CA (1997) Perceiving the causes of coarticulatory acoustic variation: consonant voicing and vowel pitch. *Percept Psychophys* 59:1141-1152.
- Pastore R, Li X, Layer J (1990) Categorical perception of nonspeech chirps and bleats. *Percept Psychophys* 48:151-156.
- Picton T, John M, Dimitrijevic A, Purcell D (2003) Human auditory steady-state responses. *Int J Audiol* 42: 177-219.
- Plomp R, Pols L, van de Geer J (1967) Dimensional Analysis of Vowel Spectra. *Journal of the Acoustical Society of America* 41:707-712.
- Plomp R, Smoorenburg GF, North Atlantic Treaty Organization. Advisory Group on Human Factors. (1970) Frequency analysis and periodicity detection in hearing. [The proceedings of the international symposium on frequency analysis and periodicity detection in hearing, held at Driebergen, the Netherlands, June [23-27, 1969]. Leiden,: Sijthoff.

- Poeppel D, Phillips C, Yellin E, Rowley H, Roberts T, Marantz A (1997) Processing of vowels in supratemporal auditory cortex. *Neurosci Lett* 221:145-148.
- Pols L, van der Kamp L, Plomp R (1969) Perceptual and physical space of vowel sounds. *J Acoust Soc Am* 46:458-467.
- Pressnitzer D, McAdams S (1999) Two phase effects in roughness perception. *J Acoust Soc Am* 105:2773-2782.
- Pressnitzer D, McAdams S, Winsberg S, Fineberg J (2000) Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. *Percept Psychophys* 62:66-80.
- R computer program . Version 2.10.1. Vienna, Austria: R Foundation for Statistical Computing; 2010.
- R computer program . Version 2.8.1. Vienna, Austria: R Foundation for Statistical Computing; 2008.
- Ragot R, Lepaul-Ercole R (1996) Brain potentials as objective indexes of auditory pitch extraction from harmonics. *Neuroreport* 7:905-909.
- Roberts T, Ferrari P, Poeppel D (1998) Latency of evoked neuromagnetic M100 reflects perceptual and acoustic stimulus attributes. *Neuroreport* 9:3265-3269.
- Roberts T, Ferrari P, Stufflebeam S, Poeppel D (2000) Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. *J Clin Neurophysiol* 17:114-129.
- Ross B, Borgmann C, Draganova R, Roberts LE, Pantev C (2000) A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude modulated tones. *J Acoust Soc Am* 108: 679-691.

- Salajegheh A, Link A, Elster C, Burghoff M, Sander T, Trahms L, Poeppel D (2004) Systematic latency variation of the auditory evoked M100: from average to single-trial data. *Neuroimage* 23:288-295.
- Sams M, Hämäläinen M, Hari R, McEvoy L (1993) Human auditory cortical mechanisms of sound lateralization: I. Interaural time differences within sound. *Hear Res* 67:89-97.
- Samson S (2003) Neuropsychological studies of musical timbre. *Ann N Y Acad Sci* 999:144-151.
- Samson S, Zatorre R (1994) Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologia* 32:231-240.
- Samson S, Zatorre R, Ramsay J (1997) Multidimensional scaling of synthetic musical timbre: perception of spectral and temporal characteristics. *Can J Exp Psychol* 51:307-315.
- Samson S, Zatorre RJ, Ramsay JO (2002) Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. *Brain* 125:511-523.
- Saupe K, Widmann A, Bendixen A, Müller MM, Schröger E (2009) Effects of intermodal attention on the auditory steady-state response and the event related potential. *Psychophysiology* 46: 321-327.
- Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. *Curr Opin Neurobiol* 15: 454-458.
- Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci* 32: 9-18.

- Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* 12: 106-113.
- Seither-Preisler A, Krumbholz K, Lütkenhöner B (2003) Sensitivity of the neuromagnetic N100m deflection to spectral bandwidth: a function of the auditory periphery? *Audiol Neurootol* 8:322-337.
- Senkowski D, Molholm S, Gomez-Ramirez M, Foxe JJ (2006) Oscillatory Beta Activity Predicts Response Speed during a Multisensory Audiovisual Reaction Time Task: A High-Density Electrical Mapping Study. *Cereb Cortex* 16: 1556-1565.
- Senkowski D, Schneider TR, Foxe JJ, Engel AK (2008) Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci* 31: 401-409.
- Senkowski D, Talsma D, Grigutsch M, Herrmann CS, Woldorff MG (2007) Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia* 45: 561-571.
- Shahin A, Roberts L, Miller L, McDonald K, Alain C (2007) Sensitivity of EEG and MEG to the N1 and P2 auditory evoked responses modulated by spectral complexity of sounds. *Brain Topogr* 20:55-61.
- Shahin A, Roberts L, Pantev C, Trainor L, Ross B (2005) Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport* 16:1781-1785.

- Shamma S, Klein D (2000) The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *J Acoust Soc Am* 107:2631-2644.
- Shepard R (1962) The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. *Psychometrika* 27:219-246.
- Simon JZ, Wang Y (2005) Fully complex magnetoencephalography. *J Neurosci Methods* 149: 64-73.
- Singh P, Bregman A (1997) The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. *J Acoust Soc Am* 102:1943-1952.
- Singh P, Hirsh I (1992) Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *J Acoust Soc Am* 92:2650-2661.
- Sinnott J, Brown C, Malik W, Kressley R (1997) A multidimensional scaling analysis of vowel discrimination in humans and monkeys. *Percept Psychophys* 59:1214-1224.
- Slawson A (1968) Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *J Acoust Soc Am* 43:87-101.
- Smith E, Lewicki M (2006) Efficient auditory coding. *Nature* 439:978-982.
- Soeta Y, Nakagawa S (2006) Complex tone processing and critical band in the human auditory cortex. *Hear Res* 222:125-132.
- Sohmer H, Pratt H, Kinarti R (1977) Sources of frequency following response (FFR) in man. *Electroencephalogr Clin Neurophysiol* 42: 656-664.
- SPSS computer program . Version 16.0. Chicago, IL: SPSS Inc.; 2007.

- Steeneken HJM, Houtgast T (1980) A physical method for measuring speech-transmission quality. *J Acoust Soc Am* 67: 318-326.
- Stein BE, Meredith MA, Huneycutt WS, McDade L (1989) Behavioral Indices of Multisensory Integration: Orientation to Visual Cues is Affected by Auditory Stimuli. *J Cogn Neurosci* 1: 12-24.
- Stevens KN (1998) *Acoustic phonetics*. Cambridge, Mass.: MIT Press.
- Stufflebeam S, Poeppel D, Rowley H, Roberts T (1998) Peri-threshold encoding of stimulus frequency and intensity in the M100 latency. *Neuroreport* 9:91-94.
- Sumbly WH, Pollack I (1954) Visual Contribution to Speech Intelligibility in Noise. *J Acoust Soc Am* 26: 212-215.
- Suzuki R, Shimodaira H (2009) pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 1.2.
- T C, Shamma S *NSL Toolbox*.
- Talsma D, Doty TJ, Strowd R, Woldorff MG (2006) Attentional capacity for processing concurrent stimuli is larger across sensory modalities than within a modality. *Psychophysiology* 43: 541-549.
- Tanji K, Leopold DA, Ye FQ, Zhu C, Malloy M, Saunders RC, Mishkin M (2010) Effect of sound intensity on tonotopic fMRI maps in the unanesthetized monkey. *Neuroimage* 49:150-157.
- Tavabi K, Obleser J, Dobel C, Pantev C (2007) Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. *Eur J Neurosci* 25:3155-3162.

- Thorpe LA, Trehub SE, Morrongiello BA, Bull D (1988) Perceptual Grouping by Infants and Preschool Children. *Developmental Psychology* 24:484-491.
- Tiitinen H, Alho K, Huotilainen M, Ilmoniemi R, Simola J, Näätänen R (1993) Tonotopic auditory cortex and the magnetoencephalographic (MEG) equivalent of the mismatch negativity. *Psychophysiology* 30:537-540.
- Tiitinen H, Mäkelä A, Mäkinen V, May P, Alku P (2004) Periodic glottal excitation and formant frequencies in the perception of vowels. *Neurol Clin Neurophysiol* 2004:103.
- Tiitinen H, Mäkelä A, Mäkinen V, May P, Alku P (2005) Disentangling the effects of phonation and articulation: hemispheric asymmetries in the auditory N1m response of the human brain. *BMC Neurosci* 6:62.
- Trainor L, Wu L, Tsang C (2004) Long-term memory for music: infants remember tempo and timbre. *Dev Sci* 7:289-296.
- Trehub SE, Endman MW, Thorpe LA (1990) Infants' perception of timbre: classification of complex tones by spectral structure. *J Exp Child Psychol* 49:300-313.
- Tucker DA, Dietrich S, McPherson DL, Salamat MT (2001) Effect of stimulus intensity level on auditory middle latency response brain maps in human adults. *J Am Acad Audiol* 12:223-232.
- Valdes JL, Perez-Abalo MC, Martin V, Savio G, Sierra C, Rodriguez E, Lins O (1997) Comparison of Statistical Indicators for the Automatic Detection of 80 Hz Auditory Steady State Responses. *Ear and Hearing* 18: 420-429.

- van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45: 598-607.
- Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, Fourth Edition. New York: Springer.
- Vihla M, Lounasmaa O, Salmelin R (2000) Cortical processing of change detection: dissociation between natural vowels and two-frequency complex tones. *Proc Natl Acad Sci U S A* 97:10590-10594.
- Vihla M, Salmelin R (2003) Hemispheric balance in processing attended and non-attended vowels and
- von Bismarck G (1974a) Timbre of Steady Sounds. *Acustica* 30:147-159.
- von Bismarck G (1974b) Sharpness as an Attribute of the Timbre of Steady Sounds. *Acustica* 30:159-172.
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13:600-612.
- Werker JF, Yeung HH (2005) Infant speech perception bootstraps word learning. *Trends Cogn Sci* 9:519-527.
- Wickham H (2009) *ggplot2: elegant graphics for data analysis*: Springer New York.
- Yau JM, Hollins M, Bensmaia SJ (2009) Textural timbre: The perception of surface microtexture depends in part on multimodal spectral cues. *Commun Integr Biol* 2:344-346.

- Yvert B, Crouzeix A, Bertrand O, Seither-Preisler A, Pantev C (2001) Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans. *Cereb Cortex* 11:411-423.
- Yvert B, Fischer C, Bertrand O, Pernier J (2005) Localization of human supratemporal auditory areas from intracerebral auditory evoked potentials using distributed source models. *Neuroimage* 28:140-153.
- Zatorre R, Bouffard M, Belin P (2004) Sensitivity to auditory object features in human temporal neocortex. *J Neurosci* 24:3637-3642.