ABSTRACT

Title of Document: ON THE WAY TO LINGUISTIC REPRESENTATION: NEUROMAGNETIC EVIDENCE OF EARLY AUDITORY ABSTRACTION IN THE PERCEPTION OF SPEECH AND PITCH

Philip J. Monahan, Ph.D., 2009

Directed By: Assoc. Prof. William J. Idsardi, Linguistics
Prof. David Poeppel, Linguistics & Biology

The goal of this dissertation is to show that even at the earliest (non-invasive) recordable stages of auditory cortical processing, we find evidence that cortex is calculating abstract representations from the acoustic signal. Looking across two distinct domains (inferential pitch perception and vowel normalization), I present evidence demonstrating that the M100, an automatic evoked neuromagnetic component that localizes to primary auditory cortex is sensitive to abstract computations. The M100 typically responds to physical properties of the stimulus in auditory and speech perception and integrates only over the first 25 to 40 ms of stimulus onset, providing a reliable dependent measure that allows us to tap into early stages of auditory cortical processing. In Chapter 2, I briefly present the episodicist position on speech perception and discuss research indicating that the strongest episodicist position is untenable. I then review findings from the mismatch negativity literature, where proposals have been made that the MMN allows access into linguistic representations supported by auditory cortex. Finally, I conclude the Chapter with a discussion of the previous findings on the M100/N1. In Chapter 3, I present neuromagnetic data showing that the response properties of the M100 are sensitive to the missing fundamental component using well-controlled stimuli. These findings suggest that listeners are reconstructing the inferred pitch by 100 ms after stimulus onset. In Chapter 4, I propose a novel formant ratio algorithm in which the *third* formant (F3) is the normalizing factor. The goal of formant ratio proposals is to provide an explicit algorithm that successfully "eliminates" speaker-dependent acoustic variation of auditory vowel tokens. Results from two MEG experiments suggest that auditory cortex is sensitive to formant ratios and that the perceptual system shows heightened sensitivity to tokens located in more densely populated regions of the vowel space. In Chapter 5, I report MEG results that suggest early auditory cortical processing is sensitive to violations of a phonological constraint on sound sequencing, suggesting that listeners make highly specific, knowledge-based predictions about rather abstract anticipated properties of the upcoming speech signal and violations of these predictions are evident in early cortical processing.

ON THE WAY TO LINGUISTIC REPRESENTATION: NEUROMAGNETIC EVIDENCE OF EARLY AUDITORY ABSTRACTION IN THE PERCEPTION OF SPEECH AND PITCH

By

Philip Joseph Monahan

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
2009

Advisory Committee:
Associate Professor William J. Idsardi, Co-Chair
Professor David Poeppel, Co-Chair
Professor Norbert Hornstein
Professor Colin Phillips
Professor Nan Bernstein Ratner, Dean's Representative

This dissertation is dedicated to

Angus (1992-2004) and Otis (1991-2005) Monahan.

Acknowledgments

I could not have asked for a more stimulating and supportive environment to complete my graduate work. The fancy equipment was nice, but it was the people that made it worthwhile. It goes without saying that my advisors William Idsardi and David Poeppel have had the largest impact on my graduate school experience. While there is a slight chance I might have been able to get through without them, it certainly would not have been as entertaining (or intellectually stimulating). Their ideas have shaped this thesis immeasurably, and they have always pushed me to consider *the big picture*. Their support, even when things were looking down, is something that I will never forget and likely can never repay.

I owe the members of my dissertation committee a debt of gratitude – Norbert Hornstein, Colin Phillips and Nan Bernstein-Ratner – because without them, I would not be writing this section. Norbert's unyielding enthusiasm for seemingly everything (linguistic and otherwise) is admirable and contagious, and his daily cookie rounds made the days that much better. Colin always asks the right questions, helping me to place my work in a larger context, refine the argumentation and question the assumptions. I also greatly appreciated the detailed and constructive comments he provided on this thesis. Finally, I'd like to thank Nan for forcing me to consider additional implications I would not normally have though about. In addition to my committee, Juan Uriagereka deserves special mention for helping with the job application process and his support.

Thanks go to Kathi Faulkingham, Robert Magee and Kimberly Kwok, who made navigating university bureaucracy so much easier, surely saving me numerous potential headaches. Additionally, Jeff Walker provided invaluable MEG lab support throughout these experiments and deserves recognition.

In addition to the exceptional faculty and staff support I received, my office-mates in Marie Mount Hall 3416F, both past and present: Pedro Alcocer, Diogo Almeida, So-One Hwang, Utako Minai, Ariane Rhone, Jon Sprouse, Yi-Min Tien and Masaya Yoshida made it truly the best office in the department over the past five years.

Ellen Lau, Yuval Marton, Rebecca McKeown, Chizuru Nakao, Eri Takahashi and Stacey Trock have been wonderful classmates. I would like to particularly thank Ellen Lau for being an outstanding colleague and co-teacher. Thanks also to my soccer buddies: Johannes Jurka and Akira Omaki. These members of the CNL lab made my life easier during my time here and deserve mention: Brian Dillon, Robert Fiorentino, Nina Kazanina, Minna Lehtonen, Hajime Ono, Leticia Pablos, Mathias Scharinger, Kevin de Souza and Matt Wagers.

Those that initially got me interested in linguistics at the University of Florida also deserve thanks: Paul Kotey, who taught my first linguistics class and Caroline Wiltshire and Eric Potsdam who supervised both my B.A. and M.A. theses, respectively. They were some of the best examples of how to be a terrific educator and an outstanding researcher. After graduating from UF, I was lucky enough to spend a year at the University of California, San Diego. There, I had a fantastic time working with

Maria Polinsky, Shin Fukuda, Nayoung Kwon and Eric Baković, and living in San Diego was pretty nice!

I would also like to thank my parents, Philip and Bernadette Monahan. I could not have asked for two more loving and caring people growing up. Thank you. Finally, I'd like to thank Eri Takahashi. She has been there with me the entire way, and made this long journey so much easier. Her love, support and devotion are more than I could ever ask and without her, this would not have been possible. I love you, and I cannot wait to begin our new life together in Spain.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A standard assumption is that brain is a computational device (Turing, 1950; Pylyshyn, 1985). Sensory receptors receive exogenous stimuli (light, sound, etc.), perform computations on this input and translate it into a representational format compatible with endogenous representations. This transduction is what allows us, as humans, to interact with and navigate our environment. The computations that perform this mapping and the stored representations that are contacted in the course of perception have typically been investigated on a modality-by-modality basis.

In the domain of speech perception, the fundamental problem is that the time-varying acoustic waveform, latent with talker- and context-dependent variation, must be mapped onto a series of stored representations (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Studdert-Kennedy, 1976, 1980; Liberman, 1996; Pisoni & Luce, 1987; Jusczyk & Luce, 2002; Diehl, Lotto, & Holt, 2004). The output of the speech perception process is a code that can make contact with the long-term memory representations. Models of speech perception have differed both on the nature of the long-term memory representation (and intermediate representations) and on the nature of the processes responsible for converting the acoustic waveform into a memory-representation. A traditional assumption from linguistics is that this code and the long-term memory representations are discrete, abstract representations devoid of the acoustic variation found in the speech signal (Jakobson, Fant, & Halle, 1952;

Chomsky & Halle, 1968). This has been the working assumption within the field of speech perception for quite some time (Studdert-Kennedy, 1976, 1980). Some models of spoken word recognition in the past couple decades, however, have proposed that the task is accomplished through the mapping of raw spectral properties (spectra, actual acoustic tokens) directly onto long-term stored memory representations. That is, there is no level of phonetics or phonology and large amounts of acoustic detail are stored in memory (Klatt, 1989; Johnson, 1997; Pisoni, 1997; Goldinger, 1998; Bybee, 2001; Pierrehumbert, 2002; Hawkins, 2003). Recent perceptual (McQueen, Cutler, & Norris, 2006), electrophysiological (Phillips, et al., 2000; Phillips, 2001; Kazanina, Phillips, & Idsardi, 2006), and neurophysiological studies (Formisano, de Martino, Bonte, & Goebel, 2008; Obleser & Eisner, 2009), however, have reinforced the existence of prelexical abstract representations in speech perception.

The goal of this dissertation is not to necessarily distinguish between these two positions, but more to show that even at the earliest (non-invasive) recordable stages of auditory processing, we find evidence that cortex is calculating abstract representations from the auditory and speech signals. In particular, looking across two distinct domains (inferential pitch perception and vowel normalization), I present evidence demonstrating that the M100, an early, automatic evoked neuromagnetic component that localizes to primary auditory cortex (Heschl's Gyrus and planum temporale), is sensitive to abstract computations over the auditory signal. The M100 (the MEG equivalent of the ERP N1) has typically shown responsiveness to physical properties of the stimulus in auditory and speech perception (Roberts, Ferrari, Stufflebeam, & Poeppel, 2000; though a handful of examples do exist that demonstrate some level of perceptual influence on the response properties of the M100) and integrates only over the first 25 ms to 40 ms of stimulus onset (Forss, Mäkelä, McEvoy, & Hari, 1993; Gage & Roberts, 2000; Gage, Roberts, & Hickok, 2006). For these reasons, the M100 provides a reliable, dependent measure that allows us to tap into the earliest stages of auditory cortical processing. In particular, I use the M100 not to study its response properties but to better understand the extent and range of computations available at these early stages of auditory cortical processing. While more work is needed to assess whether findings of this type do in fact argue against strong exemplar models, it seems that simple storage of spectra alone is not enough (cf., Klatt, 1989).

In Chapter 2, I briefly present the episodicist position on speech perception and subsequently discuss research, which suggests that the strongest episodicist position is untenable; that is, various types of data suggest that prelexical abstract representations are present and that talker-dependent acoustic information and linguistic-content are streamed early in auditory processing (consistent with the findings from Chapter 4). Then, I discuss findings from the mismatch negativity (MMN) literature, where proposals have been made that the MMN allows us to tap into the nature of the representations supported by auditory cortex and that there seems to be MMN evidence which demonstrates effects that could only be found at the level of abstract phonological representations. Finally, I wrap up the Chapter with a discussion of the previous findings on the M100/N1, since this is the dependent measure exploited in Chapters 3 and 4.

In Chapter 3, I present neuromagnetic data that suggests that the response properties of the M100 are sensitive to the missing fundamental component in auditory perception using well-controlled stimuli. Moreover, these findings demonstrate that the extraction of the missing fundamental component is achieved in the earliest stages of auditory cortical processing. Understanding the time course of how listeners reconstruct a missing fundamental component in an auditory stimulus remains elusive. Two outside tones of four-tone complex stimuli were held constant (1200 Hz and 2400 Hz), while two inside tones were systematically modulated (between 1300 Hz and 2300 Hz), such that the restored fundamental (also knows as “virtual pitch”) changed from 100 Hz to 600 Hz. Constructing the auditory stimuli in this manner controls for a number of spectral properties known to modulate the neuromagnetic signal. The tone complex stimuli only diverged on the value of the missing fundamental component. I compared the M100 latencies of these tone complexes to the M100 latencies elicited by their respective pure tone (spectral pitch) counterparts. The M100 latencies for the tone complexes matched their pure sinusoid counterparts, while also replicating the M100 temporal latency response curve found in previous studies. Our findings suggest that listeners are reconstructing the inferred pitch by roughly 100 ms after stimulus onset and are consistent with previous electrophysiological research suggesting that the inferential pitch is perceived in early auditory cortex.

In Chapter 4, I present a novel solution to the vowel normalization problem and present MEG data that suggests that auditory cortex is sensitive to the computations required by this algorithm. A long-standing question in speech perception research has been to understand how listeners extract linguistic content from a highly variable acoustic input. In the domain of vowel perception, *formant ratios*, or the calculation of relative differences between vowel formants, have been a sporadically proposed solution. I propose a novel formant ratio algorithm in which the *third* formant (F3) is taken as the normalizing factor, with the first (F1) and second (F2) formants are ratioed against it. Results from two magnetoencephelographic (MEG) experiments are presented, which suggest that auditory cortex is sensitive to formant ratios. These findings also demonstrate that the perceptual system shows heightened sensitivity to tokens located in more densely populated regions of the vowel space. Statistical evidence that this algorithm is computationally plausible in eliminating speaker-dependent variation based on age and gender from vowel productions is also presented. I conclude that these results present an impetus to reconsider formant ratios as a legitimate mechanistic component in the solution to the problem of speaker normalization.

Chapter 5 presents a slight departure from the previous two Chapters, but is still fundamentally concerned with the early stages of auditory cortical processing. Constraints on how speech sounds are sequenced in spoken language comprehension and production constitute an important part of one's phonological knowledge. Understanding how this knowledge is deployed - and how it might influence early auditory processing - is critical for determining the mechanisms involved in speech perception. I report MEG results that suggest early auditory cortical processing is sensitive to violations of a universal phonological constraint, specifically that syllable final obstruent consonant clusters must agree in voicing. By 150 ms after the second (violating) obstruent, I find a reliable difference in the areal response amplitude (RMS) of the MEG temporal waveform to syllables in which the final obstruents disagree in voicing. These findings suggest that listeners make highly specific, knowledge-based predictions about rather abstract anticipated properties of the upcoming speech signal and violations of these predictions are evident in early cortical processing.

Chapter 2: Evidence for Prelexical Abstraction

The Problem of Speech Perception

The rapidity and ease with which humans comprehend language is undeniably remarkable. Understanding speech is effortless, automatic, subconscious and uniquely human. To determine how we process and recognize spoken language is to gain insight into one profound aspect of the human mind. The mechanisms we use to accomplish speech recognition, however, remain largely unmapped. While a multitude of models and proposals regarding the processes and representations responsible for accomplishing this task have been put forward, the most fundamental set of processes underlying speech perception are simple, and I believe largely uncontroversial: perturbations of air molecules make contact with the peripheral auditory system, which sends this information to auditory cortex in the brain, whose responsibility is to use this information to make contact with some sort of memory representation. The disagreements, debates and contentions in the field of speech perception over the last 60 years have concerned the more difficult, but also more interesting questions: those pertaining to the nature of process and representation in spoken word recognition and speech perception, the relative importance of inter-talker and context dependent variability in the speech signal, whether speech perception involves a direct mapping between the input and words or whether there are intermediate sublexical representations, among others.

Models of speech perception and spoken word recognition have varied along a number of different dimensions. For example, some models explicitly use feedback as part of the perceptual process (Analysis by Synthesis: Halle & Stevens, 1959; Halle & Stevens, 1962; TRACE: McClelland & Elman, 1986), while others are strictly feed-forward (MERGE: Norris, McQueen, & Cutler, 2000). Meanwhile, some models either implicitly (McClelland & Elman, 1986; Norris, 1994) or explicitly (Jackson & Morton, 1984) assume that lexical representations are encoded in memory as abstract codes, retaining only information relevant for lexical distinctions, while other models propose that most (if not all) acoustic detail is preserved up through lexical selection (Goldinger, 1996b, 1998; Johnson, 1997). In this dissertation, I focus on the question of abstraction in speech representations and very early neurophysiological responses that seem to encode abstract properties of the stimulus.

Speech perception researchers have traditionally assumed that the auditory speech signal is recoded into increasingly more abstract representations that parallel levels of linguistic representation: (1) auditory, (2) phonetic, (3) phonological, (4) lexical, (5) syntactic and (6) semantic (Studdert-Kennedy, 1976; Pisoni & Luce, 1987; Phillips, 2001). Auditory information is recoded into a phonetic representation of the speech token. These phonetic representations include some level of linguistically relevant category distinctions, while maintaining within-category contrast ability.[1] Phonetic representations are typically gradient in nature, reflecting Gaussian distributions in a multi-dimensional phonetic space (Pierrehumbert, 2002). The phonetic

[1] Within-category discriminability is more typically found with vowels than consonants (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; see Kuhl, 2004 for a review of prototypicality and magnet effects in vowels).

representation is then converted into a more abstract, all-or-none phonological category representation, which can be manipulated by symbolic processes and does not encode information about how prototypical the particular token is, etc. Phonological category representations are typically assumed to be the basis on which lexical items are constructed. Network models such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) have included the phoneme as a level of representation mediating between the acoustic/phonetic input and words.

Marslen-Wilson & Warren (1994) argued against this traditional view that phonemic representations exist between the acoustic input and the word.[2] Instead, they proposed that listeners map directly from features onto words: "On the model we are advocating, where featural information is projected directly onto the lexical level, there will be no prelexical integration of featural cues to identify higher order units" (Marslen-Wilson & Warren, 1994, p. 655). In a series of experiments, they constructed C(C)VC stimuli that contained a subcategorical mismatch between the formant transitions out of the vowel and into the final consonant. They predicted that if a subcategorical mismatch impairs perception in the cases of words but not in the cases of non-words, this is evidence that these features are mapped directly onto lexical

[2] It should be emphasized that Marslen-Wilson & Warren (1994) do not seem to be arguing against multiple levels of representation between the input and the word per se, but instead against an explicit intermediate level of phonemic representation. The "features" they refer to appear to be discretized in nature (similar to what one might find in traditional phonological theory), and this mapping is not between the continuous acoustic space and the word-level (as one might find in a strong episodic model), but between the feature-level and the word-level. In their explicit implementation of this feature-to-word model of spoken word recognition, Gaskell & Marslen-Wilson (1997) modeled the input to the distributed neural network as binary, discrete features (in fact, a considerable subset of the features used were taken directly from Jakobson, et al., 1952).

items (featural hypothesis). Were these features mapped onto intermediate, prelexical representations, then perception should be impaired in the case of a featural mismatch regardless of whether or not the stimulus is a word (phonemic hypothesis). By cross-splicing the CV sequence *jo* (as in ‘job’, ‘jog’, etc.), taken from the words *job* and *jog* and the nonword *jod*, onto a word-final [b] (taken from a token of *job*), they created three phonemically identical but featurally distinct words. The list of conditions with examples is presented in Table 2.1.

Type of Sequence	Notation	Example
Word Sequences		
1. Word 1 + Word 1	W1W1	job + job
2. Word 2 + Word 1	W2W1	jog + job
3. Nonword 3 + Word 1	N3W1	jod + job
Nonword Sequences		
1. Nonword 1 + Nonword 1	N1N1	smob + smob
2. Word 2 + Nonword 1	W2N1	smog + smob
3. Nonword 3 + Nonword 1	N3N1	smod + smob

Table 2.1: Stimuli used in Marslen-Wilson & Warren (1994)

Recreated from Marslen-Wilson & Warren (1994, p. 657).

In two of the word cases (i.e., W2W1 and N3W1), there should be a featural mismatch between the formant transitions in the vowel and the final [b] stop-burst. At the phonemic level, all items in the Word Sequences were words. Once the vowel is processed, however, listeners might predict, given the available acoustic/featural cues, that W1W1 and W2W1 were words, while N3W1 was not. In the non-word sequences, the CCV *smo* was spliced from one word (*smog*, W2N1) and two non-words (*smob*, N1N1; *smod*, N3N1). Items in the condition W2N1 should be perceived as a *word* up to the vowel, while items in the condition N3N1 should be perceived as a *nonword* up to the vowel. Given that there is a featural mismatch in W2W1 and N3W1, but both are ultimately words, there should be an equal slowdown in the processing in either condition according to both the featural and phonemic hypotheses. The two hypotheses diverge in their predictions in the nonword cases. In particular, the phonemic hypothesis predicts that both W2N1 and N3N1 should be similarly affected by conflicting cues. If features are being mapped onto an intermediate level of representation (i.e., the phoneme), then both contain a conflict and should equally affect participants' "No" responses. The featural hypothesis, however, predicts that N3N1 should behave quite differently from W2N1. Their responses should not slowdown for N3N1 compared with N1N1 because there is no lexical entry onto which N3 could be mapped, so in some sense, it does not matter whether there is a conflicting cue or not. On the other hand, there is a clear conflict in the W2N1 case. The string W2 can be mapped onto a lexical representation and the conflicting cue in N1 should cause difficulty for participants. In a series of experiments (lexical decision, gating, phonetic decision), they found that the N1N3 condition patterned much more closely to the N1N1 condition than to the W2N1 condition, and they consistently found a reliable difference between the N3N1 and W2N1 conditions. For example, W2N1 showed much slower reaction times and higher error rates than N3N1 in a lexical decision task. They take these findings to suggest that features can be mapped directly onto words, and that an intermediate level of phonemic representation is not required (cf., TRACE: McClelland & Elman, 1986). The findings in this paper were later modeled by a distributed connectionist network (Distributed Cohort Model (DCM): Gaskell & Marslen-Wilson, 1997).

Interestingly, they consider the features implicated in this model of speech perception to be those assumed by phonological theory (see Lahiri & Marslen-Wilson, 1991). There are two potential problems with this assumption: 1) Formant transitions to place of articulation of a preceding/following consonant on the vowel have yet to be described in terms of phonological/distinctive features and 2) under current thinking in phonological theory (McCarthy, 1988; Halle, 1995), features are not independently stored units, but are instead intrinsically linked to other features in a hierarchical organization bound by a single node at the top of the representation (where the phoneme label would go). It then seems difficult to be able to access individual features without also accessing the entire phoneme (or top-node itself).[3] Subsequent psycholinguistic work, however, has shown the necessity of intervening representations between the level of the feature and the level of the word. For example, Luce & Large (2001) were able to disentangle the typically highly correlated effects of sublexical phonotactic probability (i.e., the likelihood of co-occurrence of two, three more adjacent segments) and lexical neighborhood density (i.e., the number of phonologically similar words in the lexicon). Higher phonotactic probabilities typi-

[3] One could imagine a different view of features (and their contribution to speech perception) that might be more compatible with the Marslen-Wilson & Warren (1994) data. Stevens (2002) presents a model in which the auditory perceptual system uses feature detectors to first identify major classes of speech sounds on a first pass (e.g., vowel, consonant, nasal, sonorant, etc.). On a second pass, a more detailed analysis of the signal is conducted, whereby spectral and temporal cues are analyzed to determine which vowel, or which consonant, etc. is intended.

cally lead to perceptual facilitation, while higher neighborhood densities typically lead to perceptual inhibition. By orthogonally modulating these two factors, they were able to show the facilitatory effects of phonotactic probability for both words (without the confound of neighborhood density) and nonwords and the inhibitory effects of neighborhood density (without the confound of phonotactic probability). Given that they were able to find clear facilitatory effects of phonotactic probability, particularly in the nonwords, suggests that there must be a level of representation at which these statistics can be computed (and subsequently reflected in the processing of both words and nonwords).

The notion, however, that one can map directly from the input onto words is not an uncommon one. Klatt (1989), for example, advocates a model (LAFS: Lexical Access from Spectra) of speech perception whereby "the expected spectral patterns for words […] are stored in a very large decoding network of expected sequences of spectra" (p. 192). There are no phonemes or phonetic features in this model. Instead, the computations and representations underlying speech recognition are the matching and storage of spectra. Words are selected by scoring the output of parallel matches between the input spectrum and stored spectra based on how distant they are from one another. This is done in parallel until a best match is obtained.

A similar sentiment is found in exemplar-based/episodic models of word recognition (Goldinger, 1996b, 1998) and speech perception (Johnson, 1997). What is typically considered to be noise in other models (e.g., talker variation, context-dependent variation) is instead retained and stored in the word's core long-term echoic memory representation and exploited in perception. In these types of models, each time a particular stimulus is encountered, a new stored representation is constructed (Hintzman, 1986). Listeners store details and do not recode the signal into more abstract representations, as stimulus variability is understood to be informative for perceptual processing (Pisoni, 1997). This view of speech perception is strongly inspired by Jacoby & Brooks (1984)'s work on nonanalytic concept formation and nonanalytic cognition in general and the findings from studies on categorization and memory that suggest that episodic detail is retained (Tulving & Schacter, 1990).

There seem to be two primary types of arguments one can find in support of episodic and exemplarist models in speech perception. The first type of argument arises from two types of data: One type of data suggests that listeners have more difficulty perceiving words in a list when that list is spoken by multiple talkers as opposed to just one talker. The abstractionist assumption, proponents of exemplar models claim, is that if the input is recoded into talker-independent units, then no difficulty should arise whether the list of words are being produced by one or 100 different talkers, since the stored representation of a word in the listener's lexicon is speaker-independent. The other type of data is that specific, usually talker-specific, acoustic details of previously heard spoken words either facilitate or inhibit perceptual processing of a new word. That is, participants are better (faster reaction times, higher accuracy) when they have heard the same word spoken previously by the same talker as compared to when a different talker previously produced the word. The second broad type of argument is more conceptual and less-data driven in nature: the traditional problems of speech perception (talker-normalization, invariance, etc.) are unsolvable under standard assumptions of perception (Haber, 1969), which require the elimination of variation in the signal. Consequently, in exemplar and episodic models of perception, talker-normalization, invariance and similar topics simply are not issues.

There is reasonable evidence in support of the first argument, namely that speaker-specific details can either facilitate or inhibit processing. For example, Mullennix, Pisoni, & Martin (1989) demonstrated that the intelligibility of words embedded in noise was greatly reduced when those words were spoken by more than a single talker. In the first experiment, they recorded 68 spoken words produced by 15 different talkers (both male and female). The test items were all CVC monosyllabic English words that contained a wide variety of consonants and vowels. The words were binned into high and low neighborhood densities, but were matched on word frequency and subjective familiarity. They manipulated three experimental factors: talker variability (one talker or 15 different talkers; this was a between subjects factor), lexical density (high vs. low) and signal-to-noise ratio (SNR). In the single talker group, participants heard the 68 test items all produced by the same speaker (the particular speaker was balanced across participants), while the mixed-talker group had words produced by all 15 talkers. All words were embedded in background noise and the words were presented at three different levels relative to the background noise (+10 dB, 0 dB, -10 dB).[4] The task for participants was to simply type on a computer the word they thought they heard. They found a main effect of talker-variability, where participants in the single-talker group more accurately identified the words than participants in the mixed-talker group (40.6% vs. 33.9%, respectively averaged

[4] The background noise was added to make the task more difficult for participants.

across SNRs and lexical density). As expected, they also found a main effect of SNR and lexical density. Crucially, however, Mullennix, et al. (1989) showed that switching between talkers, regardless of SNR or lexical density created more difficulty for participants than when participants heard the same list spoken by one talker. Again, the assumption from the ‘abstractionist’ literature is that there should be no difficulty if the words are being immediately converted into talker-independent representations. In a second experiment, Mullennix, et al. (1989) used the same materials and conditions (only talker-variability and lexical density in this experiment), but instead of typing the words participants thought they heard, they were required to articulate into a microphone the word they had just heard over their headset. For example, they would hear “dog” and were asked to immediately repeat aloud “dog”. In a naming task, two dependent measures can be recorded (naming latency and accuracy). They again found a reliable effect of talker-variability in both the naming latency data (single talker: 608 ms; mixed-talker: 678 ms) and the accuracy data (single talker: 95.8%; mixed-talker: 91.4%). Again, these findings demonstrate that when listeners encounter greater difficulty in a task when they are presented with multiple talkers as opposed to a single talker. How strongly these effects support an exemplar model over an abstractionist model remain to be seen, however. Abstractionist models do assume that at some point in the recognition process, talker-dependent information and linguistic content must be segregated (it is not clear that abstractionist models require talker-dependent acoustic information to be 'discarded', as is commonly assumed by proponents of exemplar models (see Pisoni, 1997)). While an extreme abstractionist model may have difficulty accounting for these findings, a model whereby both kinds of information (linguistic and speaker dependent) are analyzed, but analyzed independently should have no difficulty. It could be the case that the output of the computations required to “normalize” (i.e., segregate) the speech signal could be reinforced upon multiple exposures to the same talker and thereby facilitate the processes normalization in similar types of experiments. In other words, it should be easier to categorize the same object over and over than to categorize multiple objects.[5]

In a different series of experiments, also which purport to support episodic models (this time of the lexicon), Goldinger (1996b) demonstrated that the fine acoustic detail of words are retained in memory and consequently form the basis for lexical representations (Goldinger, 1998). In a pair of experiments, he shows that same-talker repetitions of the same word (between training and test) elicited better performance than different-talker repetitions of the same word, and that this influence could persist up to a day and even a week after training. From this data Goldinger (1996b) argues that only models which suppose that fine-acoustic detail is stored in the word’s long-term memory representation can predict this pattern of results.

The fact that speaker-dependent factors affect spoken word recognition seems apparent. The question, then, seems to not be about whether they do or do not, but what these results mean for how we concepualize the representations involved in speech perception and lexical access. Does the fact that speaker-dependent acoustic variation influences word recognition require that we give up on the notion of abstract representations and only posit exemplars in long term memory? McQueen, Cutler, &

[5] See Chapter 4 for more discussion regarding various models of speaker normalization and the algorithms proposed to eliminate/extract talker-dependent variation on vowel productions.

Norris (2006) present a perceptual learning experiment that suggests the need to posit prelexical abstractions (cf., exemplar models). Furthermore, they take a more rational approach to what these episodic findings actually suggest:

> "The critical question, however, is whether these episodic representations constitute the basic substrate of the mental lexicon or should be considered simply an adjunct to representations that are primarily abstract in nature. We argue here that evidence that listeners can show sensitivity to episodic detail should not be taken as evidence against abstract representations; further, we argue that the lexicon cannot consist solely of episodic traces" (pp. 1113-1114).

McQueen, et al. (2006) test whether listeners can make generalizations across words about sublexical properties, a task that episodic models precisely cannot do. This is because models that store only detailed acoustic representations of words in memory, they claim, cannot take advantage of sublexical regularities and generalizations. They used Dutch words that only differed in their final consonant (i.e., [f] or [s]; English: *knife*-*nice*). Participants were presented with 100 phonotactically legal words and 100 phonotactically legal nonwords in the training phase in a cross-modal priming task. The critical words ended in either [f] or [s], where the substitution of a word-final [f] for [s] or vice versa would result in a nonword. In a training phase (auditory lexical decision), one group of participants heard all the word-final [f] segments replaced by [?], which was perceptually ambiguous between [f] and [s]. The same group of participants heard [s]-final words with an unambiguous [s]. Another group of participants, however, heard the word-final [s] segments replaced by [?], and the [f]-final words with an unambiguous [f]. Therefore, one group of participants was forced to compensate for an ambiguously produced [f] and the other group was forced to compensate for an ambiguously produced [s]. The participants who heard an ambiguous [f] were biased toward categorizing the [?] as an [f] because [f] and not [s] would result in a lexical item (see Ganong, 1980). The tokens in the test phase were taken from minimal pairs, words that contrasted in their final sound ([f] vs. [s]; e.g., *doof*-*doos*) in Dutch, and were not present in the training phase. Half of the [f]-final words were paired with an ambiguous auditory prime (e.g., *doo?*-*doos*) and the other half with an unrelated prime (e.g., *krop*-*doos*). The same was true for the [s]-final words. If listeners can extend sublexical generalizations to novel lexical items (cf., exemplar models), they predict that participants familiarized with an ambiguous [?] replacing [f] (i.e., the talker produces her [f] in a strange manner) should perceive [do:?] as [do:f], and therefore be faster and more accurate to [do:f] *doof* following the prime [do:?] compared to the unrelated prime [krop]. The same group of participants should show no effect of facilitation to [do:s] *doos* following the prime [do:?] compared with the unrelated prime. And the opposite pattern of effects should hold for participants who heard an ambiguous [?] for [s] and an unambiguous word-final [f]. The found exactly what they predicted: participants given [?] for [f] in training were faster and more accurate to respond to the target [do:f] *doof* when preceded by the ambiguous prime [do:?] compared to the unrelated prime. Moreover, these participants were, in fact, slower to respond to [do:s] *doos* following the ambiguous prime [do:?] compared to [do:s] *doos* preceded by the unrelated prime. A similar pattern of results with the opposite particular findings was found for the participants exposed to [?] for [s]. Based on these findings, it seems that listeners are able generalize sublexical regularities (e.g., "I always hear [?] as [f] and never as [s]") learned from a training session and extend them to novel lexical items in such a way that they bias the perception of an ambiguous sound toward the learned generalization even though both options (assigning [?] to [s] or [f] would result in a word). Therefore, McQueen, et al. (2006) seem to have shown that participants can do exactly what exemplar models predict that they cannot do (make sublexical generalizations to novel stimuli).

Evidence for prelexical abstraction has also been provided by recent electrophysiological and neurophysiological research (see Obleser & Eisner, 2009 for a review). The hypothesis that talker-dependent factors are not 'discarded' but simply segregated early on from the linguistic content and processed independently is supported by recent fMRI results. These findings suggest distinct, though proximal, cortical areas sub-serving the extraction of linguistic content (vowel type) and talker-dependent information (Formisano, et al., 2008). They combined multivariate statistical pattern recognition with single-trial fMRI analysis and estimated the distinct activation patterns caused by individual vowel tokens. Participants heard three Dutch vowels (/i/, /a/, /u/) spoken by three different native speakers of Dutch. The speech sounds evoked areas of activation in a relatively wide portion of superior temporal cortex (consistent with Wise, et al., 1991; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; see Scott & Johnsrude, 2003; Hickok & Poeppel, 2007 for reviews), including Heschl's Gyrus and planum temporale (primary auditory regions), and also in superior temporal gyrus (STG), superior temporal sulcus (STS) and medial temporal gyrus (MTG). Using a series of machine learning classification algorithms on the fMRI data, they report that the cortical networks underlying speaker identification independent of vowel category were far more right lateralized and included the lateral part of Heschl's gyrus (Heschl's sulcus) and three regions along the anterior-posterior axis of the right STS that were adjacent to areas in vowel discrimination. The cortical maps showing sensitivity to vowel-type as opposed to speaker identity are presented in Figure 2.1.

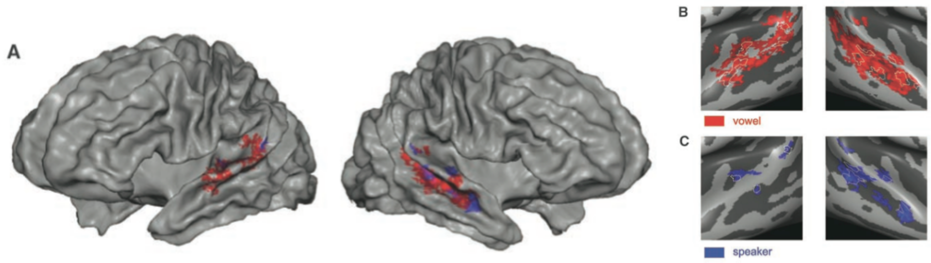


Figure 2.1: Cortical Maps corresponding to Vowel and Speaker Activity

Cortical maps illustrating the differentially active brain regions involved in vowel-type perception and speaker identity. Reprinted from Formisano, et al. (2008).

These findings suggest that at the earliest stages of auditory processing, speaker-dependent acoustic variation is segregated and processed in independent cortical regions from linguistic category information. This finding is inconsistent with models of lexical access and spoken word recognition that hypothesize that acoustic variation associated with talker-dependent factors is encoded in the core memory representation for words (Johnson, 1997; Pisoni, 1997; Goldinger, 1998).

In short, it appears that while talker-dependent effects can influence spoken word recognition (Goldinger, 1996b, 1998; Mullennix, et al., 1989; Johnson, 1997; Pisoni, 1997), there is also ample evidence that spoken word recognition and speech perception involves prelexical abstraction (McQueen, et al., 2006; Obleser & Eisner, 2009 and references therein). Given these results, it seems that the characterization of the extreme abstractionist model (all unwanted variation is discarded and unprocessed) is untenable,[6] while a strictly exemplar/episodic model (no abstraction) is also inadequate. McQueen, et al. (2006) envision a model whereby lexical representations could consist of core abstraction units augmented by a separate store of talker-dependent factors, frequency, etc. In the next section, I review electrophysiological evidence from MMN/MMF and N1/M100 studies. Results in the MMN/MMF paradigm seem to be getting us closer to a real understanding of the types of abstract representations auditory cortex can support. In Chapters 3 and 4, I present data which suggests that the N1/M100, normally only responsive to physical attributes of the stimulus, also indexes abstract computations.

Electrophysiological Evidence for Abstraction

Electrophysiology (EEG, MEG) has proven to be an exceptionally useful tool for understanding the nature of auditory and speech representations. The early electrophysiological evoked components commonly associated with auditory and speech perception (N1/N1m/M100, N1-P2 complex, MMN/MMNm/MMF) are pre-attentive and do

[6] I am unaware of proponents of this strong hypothesis, but this is the characterization of abstractionists typically found in the episodic literature.

not require a task, ideally providing researchers with a task-independent probe into the early stages of processing. The automaticity of these components, combined with the techniques' excellent, millisecond-by-millisecond temporal resolution, makes using such methods extremely powerful in understanding the nature of linguistic and auditory representations and processes employed and entertained prior to contact with a lexical-level of representation. In this section of the dissertation, I focus primarily on two components. First, I provide a brief overview of the Mismatch Negativity paradigm and response (MMN). An enormous amount of work has been done on not only understanding the response properties and the neurophysiological and cognitive sources of the MMN (see Näätänen, 1992; Näätänen, Paavilainen, Rinne, & Alho, 2007 for extensive reviews), but also the nature of auditory and linguistic representations as indexed by the MMN (Näätänen, 2001). In particular, MMN studies of speech perception have provided neurophysiological evidence for the existence of representations at the level of abstract phonology (e.g., phonemes, distinctive features) and shown sensitivity of native language phonological inventories (Näätänen, et al., 1997; István Winkler, et al., 1999; among others) and syllabic constraints (Dehaene-Lambertz, Dupoux, & Gout, 2000). This data is reviewed below. Given the nature of the MMN paradigm, however, I suggest that this particular methodology is better suited for studying the nature of representations and less well suited for understanding the set of computational processes employed to map the time-varying acoustic signal onto linguistic units. Subsequently, I briefly review the literature on the N1/N1m/M100, demonstrating that to this point, it has been thought to faithfully reflect physical properties of the stimulus. In Chapters 3 and 4, I present a series of experiments that suggest the M100, and more importantly, the neurobiological generators of the M100 can also reflect computations performed over the physical stimulus. That is, abstraction by 100 ms.

The Mismatch Negativity

The Mismatch Negativity (MMN; Mismatch Magnetic Field (MMF/MMNm) in MEG) is an electrophysiological component observed when there is “virtually any *discriminable physical change* in an otherwise repeated auditory stimulus (Gomes, Ritter, & Vaughan, 1995, p. 81)”. In this discussion, I focus on auditory elicitations of the MMN, though it should be noted that the MMN has also been observed in the visual (Tales, Newton, Troscianko, & Butler, 1999; Maekawa, et al., 2005), olfactory (Krauel, Schott, Sojka, Pause, & Fersti, 1999; Pause & Krauel, 2000) and somatosensory (Kekoni, et al., 1997; Shinozaki, Yabe, Sutoh, Hiruma, & Kaneko, 1998) modalities. In a typical MMN/MMF paradigm, participants are presented with a series of *standard* tokens interrupted by a *deviant*, which differs from the standard along some physical (or linguistic) dimension. A schematic of the trial structure for an MMN oddball design is presented in Figure 2.2.

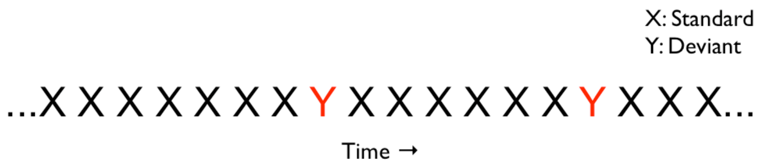


Figure 2.2: Trial structure of a Mismatch Negativity (MMN) oddball design.

Participants are presented with a series of ‘standards’ (X) occasionally interrupted by a ‘deviant’ oddball stimulus (Y), which in EEG elicits a larger negative going wave compared with the response to the standard (in MEG elicits a stronger magnetic field (larger amplitude) in response to the deviant compared with the standard). Participants are typically instructed to passively attend to the experimental stimuli.

Modulations of spectral properties of an auditory stimulus, such as the frequency, intensity and duration (Sams, Paavilainen, Alho, & Näätänen, 1985; István Winkler, et al., 1990; Gomes, et al., 1995; see Näätänen, 1992 for a review) have all been reported to reliably elicit an MMN/MMF. The probability of hearing a deviant within a given experimental block is usually on the order of 10% - 20% (sometimes < 10%). If the deviant is perceived as being perceptually distinct from the standards (the requisite physical difference between the standard and deviant is typically commensurate with behavioral discrimination thresholds for particular stimulus attributes (Näätänen, et al., 2007)), then a large, negative-going waveform (in EEG; a larger magnetic field strength in MEG) in comparison with the electrophysiological response to the standard is observed approximately 150-300 ms post-onset of the deviant stimulus. The magnitude of the MMN elicited by the deviant is determined by subtracting the grand average waveform of the electrophysiological response to the standard from the grand average waveform of the electrophysiological response to the deviant (Näätänen, 1992, 2001; Näätänen, et al., 2007). Based on recordings from monkey (Javitt, Schroeder, Steinschneider, Arezzo, & Vaughan, 1992; Javitt, Steinschneider, Schroder, Vaughan, & Arezzo, 1994), cat (Csépe, Karmos, & Molnár, 1987), MEG recordings in humans (Hari, et al., 1984) and source modeling of the EEG signal (Scherg, Vajsar, & Picton, 1989; Alain, Cortese, & Picton, 1998; see Näätänen & Alho, 1995 for a review), the neural generators of the MMN/MMF component are located in the superior temporal plane in primary (or immediately adjacent to primary) auditory cortex, roughly 3-10 mm more anterior than the source of the N1m (M100; Näätänen, et al., 2007).

The MMN/MMF is an automatic electrophysiological response. Its elicitation does not require attention on the part of the participant or active control modules. Instead, it can be elicited during sleep(Atienza, Cantero, & Dominguez-Marin, 2002; Csépe, 1995; Csépe, et al., 1987; Nashida, et al., 2000; Sculthorpe, Ouellet, & Campbell, 2009), comatose (Fischer, Morlet, & Giard, 2000; Vanhaudenhuyse, Laureys, & Perrin, 2008) and anesthetized states (Csépe, 1995; though see Simpson, et al., 2002 for limitations on the elicitation of the MMN during anaesthesia), and the magnitude and presence of the response are generally unaffected by attentional or task demands required of participants during the experimental procedure (Näätänen, 1992; Ritter, et al., 1992; Ritter & Ruchkin, 1992; Alho, et al., 1998). Therefore, during the experiment, participants are often asked to read texts (e.g., H. Tiitinen, et al., 1993) or view movies (e.g., Tervaniemi, et al., 1999; Gaeta, Friedman, Ritter, & Cheng, 2001) while passively attending to auditory stimuli.

Low-Level Auditory Elicitation of the MMN

To take an extremely straightforward example to demonstrate the phenomenon of the MMN, Sams, et al. (1985) presented participants with a series of identical 1 KHz pure sinusoids: the *standards*. Twenty-percent of the time, however, participants heard a pure sinusoid of a different frequency (1004 Hz, 1008 Hz, 1016 Hz or 1032 Hz) interspersed between the standards: the *deviant*. There were four blocks and each block had one, and only one deviant sinusoid. While participants passively listened to this series of standard and deviant auditory stimuli, electrical potentials were recorded from electrodes placed on the scalp. They observed a large negative-going deflection in the ERP waveform to the 1016 Hz and 1032 Hz deviants compared to the averaged 1 KHz standard. An extremely small, but still present, MMN was observed to the 1008 Hz sinusoid in comparison with the response to the 1 KHz standard, an auditory difference just above participants' threshold in an independent auditory discrimination task.

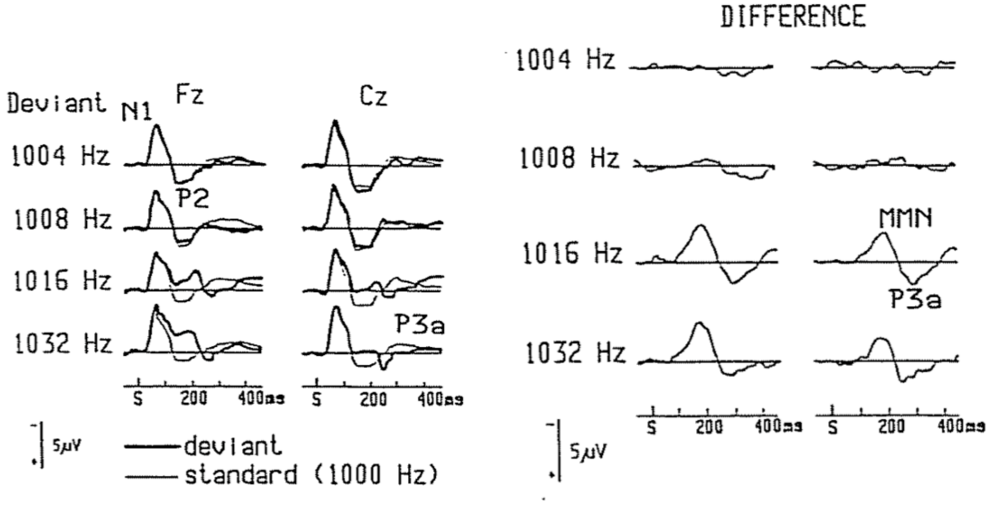


Figure 2.3: Electrophysiological responses to standard and deviant stimuli in Sams, et al. (1985).

Left panel is overlay of ERP response to standard compared with deviant per condition. Right panel is the subtraction of the ERP response to the standard compared with the ERP response to the deviant (difference wave). Notice the strong MMN in the 1016 Hz and 1032 Hz conditions (and the marginal MMN response in the 1008 Hz condition).

The sub-threshold 4 Hz difference between the deviant and standard in the 1004 Hz condition was insufficient to elicit an MMN. Thus, the detectable physical changes in the deviant (i.e., 1016 Hz, 1032 Hz, and to a lesser extent, 1008 Hz) compared to the

1 KHz standard elicited an MMN, while the undetectable perceptual differences (i.e., 1004 Hz) did not.

Provided the hypothesis that the cognitive mechanisms underlying the MMN (change detector, attentional switch, etc.) are implicated in normal, everyday perception and not confined to laboratory induced situations (Näätänen, Sams, & Alho, 1986), Winkler, et al. (1990) questioned whether a series of repeating, identical standards were sufficient to reliably infer anything about the perceptual system on a more general scale. This is because, rarely in the natural world, are humans presented with a series of truly identical (on the physical level) perceptual objects. Therefore, they asked whether an MMN could be elicited if variation was introduced into the series of standards. Given that stimulus intensity and frequency, to that point, had been the most commonly studied auditory modulations, Winkler, et al. (1990) tested both an intensity deviant and a frequency deviant while measuring evoked related potentials (ERPs) from an electrode-array placed on the scalp. The spectral frequency of the standard was 600 Hz across all conditions. A series of "substandards", as they referred to them, that varied in their stimulus intensity were synthesized. The mean intensity was 80 dB SPL. The experiment consisted of five distinct blocks, and each block had a different intensity range for the standards (i.e., 80 ± 0, 0.8, 1.6, 3.2 and 6.4 dB), and in each block, the difference in intensity between any two consecutive standards was constrained (0, 0.2, 0.4, 0.8 and 1.6 dB, respectively). The deviants in the *intensity* condition had an intensity of 70 dB SPL (600 Hz carrier frequency), and the deviants in the *frequency* condition had a spectral frequency of 650 Hz (80 dB intensity, same as the mean intensity of the standard tokens). If an MMN can be elicited despite variation in the physical properties of the standards, then an MMN should be elicited to the deviant stimuli (in both the intensity and frequency conditions) not only in the block where there was no intensity variation (80 ± 0 dB), but also in the blocks where there was variation in the intensity of the standards (i.e., 80 ± 0.8, 1.6, 3.2 and 6.4 dB). In the intensity condition (deviant: 600 Hz; amplitude: 70 dB SPL), an MMN was elicited in all blocks, including the block with the greatest intensity range (± 6.4 dB), although the peak amplitude, difference area, and end latency decreased as the range increased. In the frequency condition (deviant: 650 Hz; amplitude: 80 dB SPL), again an MMN was elicited in all blocks, including the block with the greatest intensity range (± 6.4 dB), and furthermore, the MMN peak amplitude and area decreased as the variation increased. In short, despite significant variation in the standards, an MMN was still elicited, suggesting that a static series of physically identical standards is not necessary to find an MMN.

Näätänen (1992, p. 139) proposes two plausible explanations for the elicitation of the MMN to a deviant stimulus from a series of standards:

1. *New Afferent Elements*: The neurobiological source of the MMN is the cortical response of new afferent neuronal populations. Given the long ISIs between deviant stimuli, neurons that respond to the deviant "remain active" and fire strongly when a deviant stimulus is processed. Given the short ISIs between standards, the responsible neuronal populations become saturated and strongly refractory, and therefore do not fire as strongly.

2. *Memory Trace*: The MMN results from a neuronal "code" that calculates the difference between the standard and deviant. As Näätänen (1992) notes, this implies that a stored representation of the standard must have been constructed in memory.

Compelling evidence in favor of the second interpretation arises from the finding of MMNs to omitted stimuli. Yabe, Tervaniemi, Reinikainen, & Näätänen (1997) presented participants with a series of 1 KHz standard sinusoids at regular intervals. A different inter-stimulus interval (ISI; 100, 125, 150, 200, 250, 300, 350 ms) was used in each experimental block. They found a reliable MMN to the omitted stimuli only at the very short ISIs (100 and 125 ms). Given that the MMN was not elicited by a physical signal, these findings suggest that the neurophysiological basis of the MMN cannot solely be attributed to new afferent neuronal populations responding preferentially to the deviant. Instead, these findings support the view that the MMN reflects a comparison metric of stimulus change between the computed standard and deviant. This conclusion has prompted researchers to begin to probe the nature of the constructed representation of the standard.

Returning to the Winkler, et al. (1990) experiment, they also report considerably smaller, though still present, MMNs to the substandards, particularly in the experimental blocks with the greatest amount of variation in the intensity (e.g., 80 ± 6.4 dB). They provide two alternative hypotheses regarding the nature of the constructed *standard* representations. The fundamental difference between these two hypotheses is the number of memory traces stored. One possibility is that participants construct one trace in memory that essentially a merging ("melted") representation of all the standard tokens. The more variation in the standards there is, the more "shallow" the memory trace is along the relevant physical dimension (e.g., intensity, frequency, etc.). A mismatch occurs when the deviant differs from the center of the distribution. The second alternative is that participants store each individual token of the standard independently from one another and a comparison between the deviant and each of the individual standards is made in parallel. The first alternative, namely that participants construct an *averaged* representation of the standard, was later eliminated as plausible hypothesis (István Winkler, Paavilainen, & Näätänen, 1992). Therefore, it is suggested that participants are storing each standard independently of one another. Gomes, et al. (1995) noted that in the Winkler, et al. (1990) experiment, it might have been possible for participants to keep individual substandard traces in memory and that these substandards could be reinforced, given that there was roughly 5 sec between any given substandard, which is well within the duration of time (10 sec maximum; though it appears that "dormant" standards may become reactivated (Ritter, Gomes, Cowan, Sussman, & Vaughan, 1998)) that a memory trace can persist (Böttcher-Gandor & Ullsperger, 1992; Sams, Hari, Rif, & Knuutila, 1993). Moreover, it has been shown that more than two stimulus repetitions of the standard are required to elicit an MMN (Cowan, Winkler, Teder, & Näätänen, 1993). Therefore, Gomes, et al. (1995) designed their experiment in such a way that individual discrete substandards would not be able to persist and be reinforced. Their hypothesis was that if the participants were simply storing the relevant property (feature) that was common to all the standards and not storing them as discrete individual memory traces, an MMN should still be evoked even if they do not perceive the same physical stimulus within the 10 sec window necessary to elicit an MMN. The primary physical difference between the standards and deviants was stimulus duration. The standards were all 100 ms in duration, while the deviants were all 170 ms in duration. In the final condition of the experiment, Gomes, et al. (1995) presented participants with a broad range of frequency and intensity values in the auditory stimuli. Participants heard 10 distinct frequencies (700 to 2050 Hz in increments of 150 Hz) and 10 distinct amplitudes (60 to 87 dB SPL). A given combination of intensity and frequency values would occur on average every 72 presentations. Ten percent of the trials were the deviant stimulus (170 ms in duration), which was simply a random combination of frequency and intensity from the values used for the standards. Reporting ERP data from 10 participants, they found an MMN to the deviant stimulus despite the significant variation in the standards (the only physical property each successive standard shared was that they were 100 ms in duration). Gomes, et al. (1995) go on to conclude that these data support the hypothesis that participants can extract the relevant feature shared (if there is one) by quite disparate standards and compare the relevant feature on the *deviant* token against the stored relevant feature on the standard. They are cautious to eliminate the possibility that gestalt representations of the standards can be constructed under less disparate experimental conditions, and acknowledge that this may, in fact, continue to be the case in some previous studies (István Winkler, et al., 1990). But they do conclude that under some circumstances, participants may be able to simply extract the relevant feature common to all standards (in their case, stimulus duration), and compare the value of the extracted feature against the potentially deviant stimulus.

Korzyukov, Winkler, Gumenyuk, & Alho (2003) demonstrated that abstract auditory regularities are processed in auditory cortex, and these abstract regularities are indexed by the MMN response. Korzyukov, et al. (2003) synthesized a series of pure sinusoid tones that were 30 ms in duration and ranged from 500 Hz to 1104 Hz in equal 12% steps. The tones were presented in pairs, with 120 ms ISI and 1 sec between tone pairs. The tone pairs were always one frequency step away from each other. They had a total of three experiments, but only two are relevant for the discussion at hand. In the Equiprobable experiment, the likelihood that the second tone in the pair would be a lower frequency than the first tone in the pair ('descending pitch pair') was equal to the likelihood that the second tone would be higher in frequency than the first tone in the pair ('ascending pitch pair'). In the Oddball condition, the probability of the 'ascending pitch pair' was 90%, while the 'descending pitch pair' was 10%. It should be pointed out that the sinusoids in the pairs were individual, discrete tones. The trial structures in the Korzyukov, et al. (2003) experiment are presented in Figure 2.4.

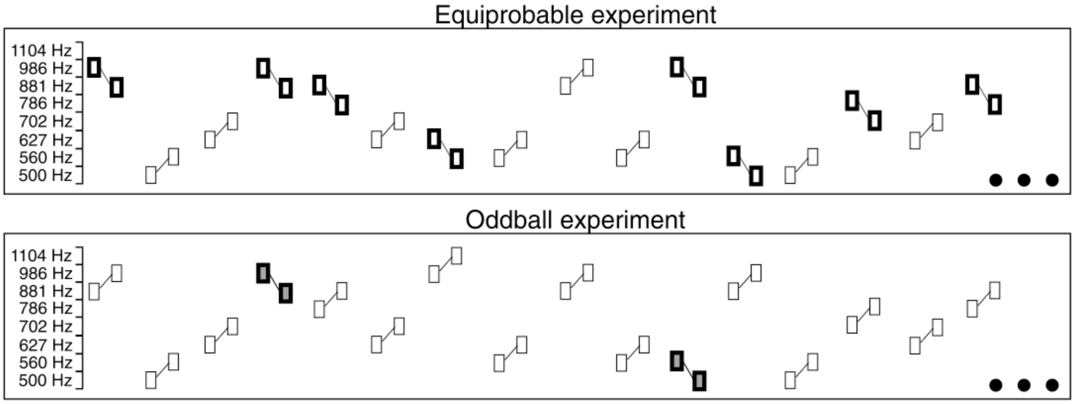


Figure 2.4: Trial structures of the Equiprobable and Oddball Experiments

The connecting lines are for illustrative purposes only. The primary difference between the two experiments was the likelihood of the 'descending pitch pairs' relative to the 'ascending pitch pairs' (50%-50% in the Equiprobable experiment; 10%-90% in the Oddball Experiment). Reprinted from Korzyukov, et al. (2003).

It is important to note that the ascending and descending pitch are inferred representations from the pattern of sinusoids and are not physically present in the stimulus. The close temporal proximity of the two sinusoids at different frequencies creates the illusory perception of a rising or falling pitch. They conceptualize these illusory pitch patterns as the construction (and breaking in the deviant case in the Oddball experiment) of abstract relational rules between the two members of the tone pair. As predicted, the 'descending pitch pairs' elicited an MMN in the Oddball experiment but not in the Equiprobable experiment.

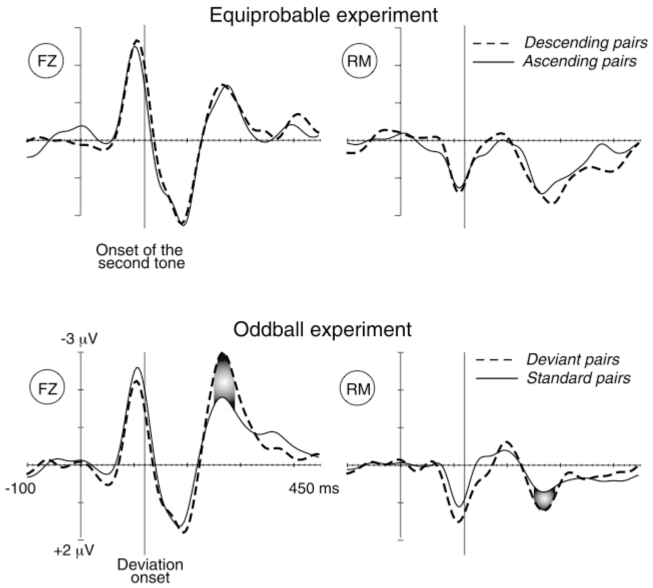


Figure 2.5: MMN waveforms to Pitch Pairs

Comparison of the electrophysiological response to the standard and deviant in the Equiprobable and Oddball experiments. Notice the MMN is elicited in the Oddball experiment, but not in the Equiprobable experiment. Reprinted from Korzyukov, et al. (2003).

In the Oddball experiment, participants likely constructed a relational rule between the tone pairs that required an ascending relationship between the pitch pair. The 10% of the time they perceived a descending relationship between the pair, the rule they formulated would be broken, and consequently, an MMN would be elicited. Crucially, however, these results suggest that the MMN can index violations of perceptual representations, despite the fact that no physical pitch contour exists in the stimuli in the experiments.

In short, an MMN is found when listeners are able to perceptually group the standards (i.e., process them as being equivalent to one another) along some meaningful dimension at some level of representation. The real interest in the MMN, then, is not what sort of stimuli can act as a deviant, but instead the extent to which participants can group different tokens of the standards into a coherent representation. In the next section, I review research that used the MMN to understand how listeners perceive speech. In particular, much of the work has focused on two, not unrelated questions: 1) the level of linguistic representation at which participants construct memory traces of the standard (phonetic vs. phonological), and thereby assessing the existence of that level of representation in speech perception, and 2) the role of native language inventory (and structural constraints) on the perception of speech sounds. Crucially, for my discussion at discussion at large, the MMN findings with respect to both of these more general questions at large have suggested that speech sound categories are processed into abstract representations and that our perception of speech is shaped by higher-order knowledge of what our native language allows and does not allow (both in terms of inventory and structure).

The MMN as an Index of Abstraction in Speech Perception

The Mismatch Negativity electrophysiological component has been exploited extensively to probe the nature of speech representations (Phillips, et al., 2000; Eulitz & Lahiri, 2004; Yeung & Phillips, 2004; Kazanina, et al., 2006) and the impact of native language phonology on the perception of speech (Näätänen, et al., 1997; István Winkler, et al., 1999; Dehaene-Lambertz, et al., 2000; Mitterer & Blomert, 2003). Research using the MMN in speech perception has demonstrated that this component reflects much higher levels of representation and processes than what can be inferred from the physical/acoustic attributes of the signal alone (see Phillips, 2001 for a review).

Näätänen et al. (1997) assessed the extent to which one's native language vowel inventory affects elicitation of the MMN, and more generally, the early stages of perceptual processing of speech. They tested native speakers of Finnish and Estonian, two closely related languages with nearly identical vowel inventories. The primary difference between the two languages' vowel inventories is that Estonian contains the vowel /õ/, which Finnish does not. Näätänen et al. (1997) synthesized vowel tokens corresponding to /e/, /ö/, /õ/ and /o/. The fundamental acoustic difference between these vowel categories is their representative value for the second formant (F2). The semi-synthetic tokens were matched on their fundamental frequency (f0) and first (F1), third (F3) and fourth (F4) formants. In a behavioral pretest, both groups of participants (Finnish and Estonians) were asked to evaluate how good of a prototype the various vowel tokens. It was clear from the responses that there was a straightforward three-way categorization for the Finnish participants and a four-way categorization for the Estonian participants. The Finnish participants did not consider the foreign phoneme (i.e., /õ/) as being prototypical of any vowel category, as they have had no prior linguistic experience with it. This is illustrated in Figure 2.6.

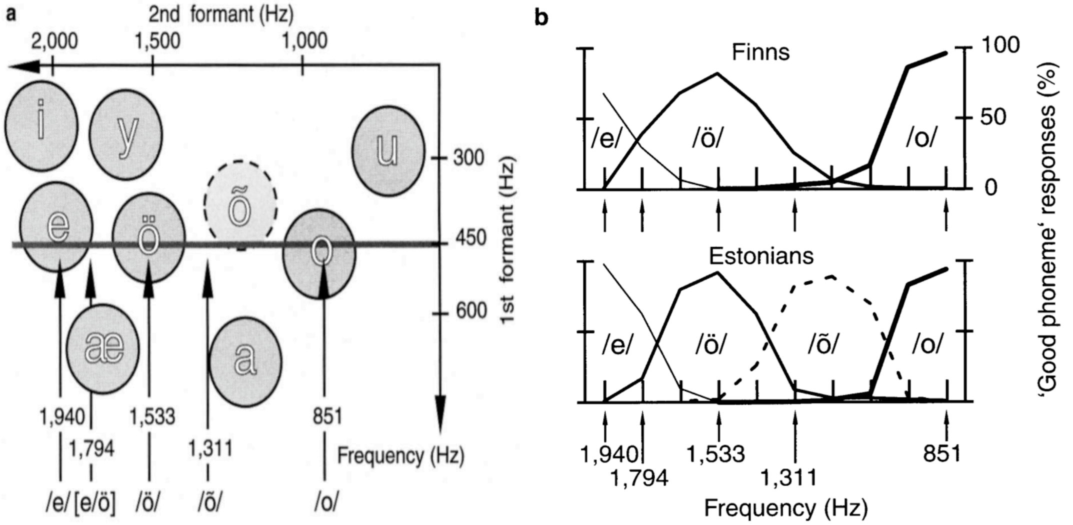


Figure 2.6: Vowel Inventory and Prototype Responses

(A) Vowel inventories of Finnish and Estonian. Estonian contains the vowel /õ/, which Finnish does not. Also evident from (A) is that the primary acoustic dimension on which /e/, /ö/, /õ/ and /o/ differ is F2. (B) 'Good phoneme' responses by the Finnish and Estonian participants for the different F2 values of the synthetic tokens. The Finnish participants identified three prototypes, while the Estonian participants identified four prototypes from the synthesized continuum. Reproduced from Näätänen et al. (1997).

In the MMN paradigm, there were two primary comparisons between the groups: 1) the response to the deviant synthetic vowel tokens (i.e., /ö/, /õ/, /o/) from the prototype standard /e/ synthetic vowel token and 2) the response to deviant sinusoids of the same frequencies of the prototype F2 values for the vowels /ö/, /õ/ and /o/ from the F2 value of the standard /e/. Given the findings from Tiitinen, May, Reinikainen, & Näätänen (1994), who found monotonic relationships between properties of the MMN and the distance of difference between the standard and deviant tokens, Näätänen et al. (1997) expected a monotonic rise in the magnitude of the MMN the further away the F2 value of the deviant was from the F2 value of the standard. Therefore, /o/ should elicit the largest MMN (because its F2 value is furthest from the standard's F2 value), while /ö/ should elicit the smallest MMN (because its F2 value is closest to the standard's F2 value) compared with the electrophysiological response to the standard /e/. The critical comparison is the response to the deviant /õ/. An MMN magnitude mid-way between /ö/ and /o/ and should be elicited in the Estonians and not in the Finnish if the MMN is reflecting phonetic/phonemic processing. If the MMN is simply reflecting the acoustic differences in the vowel tokens, then a monotonic rise in the magnitude of the MMN across all three deviants should be elicited in both groups of participants. In the vowel portion of the experiment, they found a monotonic increase across the three vowel tokens for the Estonians but not for the Finnish speakers. The difference waveforms from across the two groups are presented in Figure 2.7.

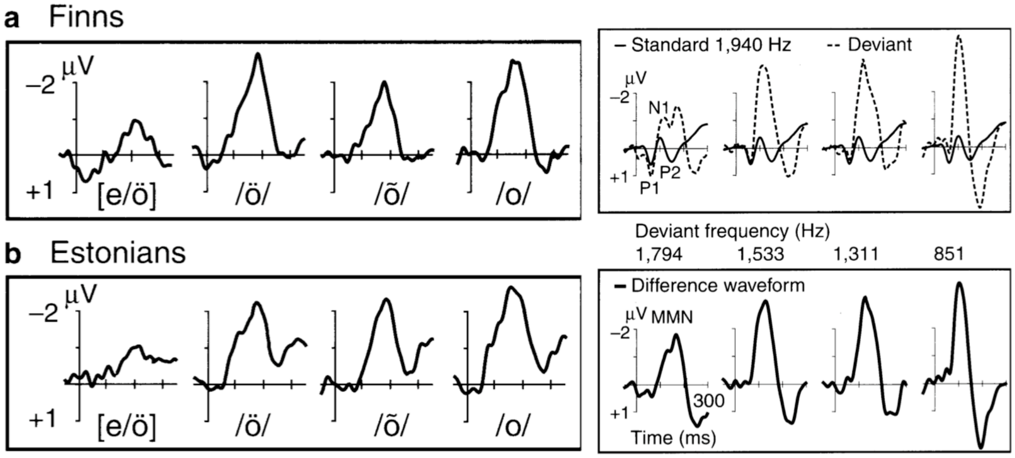


Figure 2.7: Comparison of MMN Responses between Finns and Estonians

The left panel is the difference waveforms between the standard /e/ and deviant vowel tokens. Notice that the magnitude of the MMN for /õ/ for the Estonian participants is mid-way between the MMN response to /ö/ and /o/. The magnitude of the MMN to /õ/ for the Finnish participants is reduced, however, compared to the magnitude of the MMN for /ö/ and /o/. The right panel (raw MEG waveforms on the top; difference waveforms on the bottom) is the MMN response for the Finnish participants to the pure sinusoids of the same values of F2 for the vowel tokens in the vowel-portion of the experiment. Reproduced from Näätänen et al. (1997).

In a second portion of the experiment, Finnish and Estonian participants heard pure sinusoids of the same frequencies as the F2 values for the synthetic vowel tokens. If the results from the vowel-portion of the experiment arise of the phonemic inventories of the two languages and not from the acoustic differences, then we expect to find a monotonic rise in the magnitude of the MMN across the three vowel tokens in both groups of participants. That is, in fact, what Näätänen et al. (1997) found. In both the Finnish and Estonian participants, they report a monotonic increase in the magnitude of the MMN in the predicted direction given the physical differences between the standard and deviant. This provides additional evidence that the difference in the magnitude of the MMN to /õ/ between the Finnish and Estonian participants is not solely attributable to the physical differences between the stimuli, but must also be a function of the phonemic vowel inventories of the two groups of participants. For the vowel tokens, this non-monotonic rise was evident in the amplitude of the MMN (as discussed above) and also in the Equivalent Current Dipole (ECD) strength for the respective vowel tokens (the Estonians did show a monotonic increase in the ECD strength in the predicted direction). In all, the data from Näätänen et al. (1997) seem to suggest, more narrowly, that the MMN is sensitive to the phonemic inventory of the participants' native language, and more generally, participants' native language influences low level perception of the speech signal. There are a handful of concerns with this experiment, however. First, because of the nature of the behavior pretest, we are unsure how the Finnish participants perceived tokens of

/õ/. Apparently, they did not greatly expand the distribution for /ö/ or /o/ to compensate for these intermediate tokens (though it appears from the Figure 2.6 that the distribution of /ö/ is more diffuse for the Finns as compared with the Estonians). Finnish participants were given three buttons to press during the task, while Estonian participants were given four responses. Second, with respect to the MMN responses, it is not that the Finnish participants failed to elicit an MMN to /õ/. It is simply the case that the magnitude of the MMN to /õ/ is somewhat reduced for the Finnish participants compared with the Estonian participants. One would not expect to find an MMN to /õ/ for the Finnish participants (this prediction is borne out in some experiments discussed below) if the elicitation of the MMN were solely dependent upon the phonemic inventory of the native language in this experiment.[7] Third, they do not report a difference between the MMN to /ö/ and the MMN to /o/ for the Finnish participants (in Figure 2.7 they appear to be roughly equivalent, though a more noticeable difference is evident in their Figure 4). If the pattern of effects we find are expected to be a monotonic rise dependent upon the physical distance from the standard, then we expect the Finns to also show a reliable difference between the magnitude of the MMN to /ö/ as compared with /o/. Fourth, and finally, the task at hand does not force participants to perceptually group vowel tokens into higher-order representations. Therefore, we cannot be certain that these results are attributable to a phonemic level of representation, as they could be encoding the standard in a representation that is

[7] On some level, the MMN in Näätänen et al. (1997) cannot be solely attributed to phonemic status. They predicted going into the experiment that they would find a monotonic increase in the magnitude of the MMN as a function of the physical distance between the standard and deviant. Therefore, a complete explanation of these MMN findings must attribute the incongruous findings between groups to both physical as well as linguistic differences.

quite sympathetic with its physical attributes. The reason for this is because they only used one acoustic token for each vowel type in the MMN portion of the experiment.

Winkler, et al. (1999) followed up on these results for two reasons. First, there was no within-category contrast in the Näätänen et al. (1997) experiment, therefore making it difficult to fully attribute their results to a phonetic/phonemic level of processing. Second, the Estonians were not tested on a Finnish-only vowel category, making the design unbalanced. Therefore, they compared participants from Finnish and Hungarian, another closely-related language, with contrasts that were within-category in one language and across-category in the other. The range of the Finnish vowel /e/ occupies portions of the four-dimensional (F1-F4) vowel space occupied by /é/ and /ɛ/ in Hungarian. Meanwhile, the Finnish vowels /e/ and /æ/ are located in the region of vowel space occupied by /ɛ/ in Hungarian. Consequently, they synthesized a pair of vowel tokens that would be perceived as /é/ and /ɛ/ by Hungarian participants, but only as /e/ by Finnish participants. They also synthesized a pair of vowel tokens that would be perceived as /e/ and /æ/ by Finnish participants, but only as /ɛ/ by Hungarian participants. Winkler, et al. (1999) report larger an MMN in all comparisons across groups. The MMNs in the across-category conditions, however, were significantly larger than the within-category condition in both groups of participants. Given that they had a fully crossed design and made within- and across-category contrasts, where the across-category contrasts elicited a larger MMN than the within-category contrast suggests that the MMN can be based on phonetic category representations. Both Näätänen et al. (1997) and Winkler, et al. (1999) used isolated vowels to assess the influence of native language phonemic inventories on the MMN and its influence on lower levels of speech perception.

More recently, Hacquard, Walter, & Marantz (2007) exploited the MMF to investigate the role of vowel inventory and size on the perception of vowels. The size of a vowel inventory within a given language influences the acoustic consequences of articulation. In particular, it has been reported that languages with larger vowel inventories also tend to have a larger acoustic vowel space relative to languages with smaller vowel inventories (Bradlow, 1995). To understand the influence of vowel inventory size and organization on perception, Hacquard, et al. (2007) tested native speakers of Spanish and French in an oddball MMF paradigm. The vowel spaces of Spanish and French differ on both their size and organization: 1) Spanish is a five-vowel system, while French has 12 vowels (including the five vowels found in Spanish) and 2) French has a series of vowels intervening in F2/F1 space between the vowels of Spanish (e.g., /ɛ/ intervenes between /e/ and /a/; /ɔ/ intervenes between /o/ and /a/). Participants listened to two blocks: in the first block, the standard was /o/ and the deviants were /a/ and /u/, and in the second block, the standard was /e/ and the deviants were /a/ and /ɛ/. Deviants occurred in 12.5% of the experimental trials. They hypothesized that if inventory organization had an effect on perception, then the effect of the MMF should be roughly equivalent for the /o/-/u/ pair as the /o/-/a/ pair. That is because these two pairs are equally distant in terms of intervening vowels in Spanish. In French, however, since /ɔ/ intervenes between /o/ and /a/, they predicted the MMF to be larger in the /o/-/a/ pair than in the /o/-/u/ pair if inventory size played a role. Analyzing across languages, they did not predict a difference in the magnitude of the MMF for the /o/-/u/ pair between the French and Spanish pair, while they did expect a difference between the /o/-/a/ pair, since French has an intervening vowel category there (and consequently, they expected a larger MMF for the French participants in inventory organization played a role in perception). If inventory size was the primary factor driving the MMF response, then they expected the French participants to show a larger MMF across the board, since French has the larger vowel inventory. What the found was that, across the board, a larger MMF was elicited in French participants, suggesting that vowel inventory size (the number of vowel categories in the language) and not inventory organization affects the perception of vowels. Specifically, they found a larger MMF for all vowel comparisons in the French participants compared with the Spanish participants except in the /ɛ/-/e/ pair. They took this particular pattern of results to support a model of the expanding vowel space whereby point vowels (e.g., vowels on the edge of the vowel space) are produced more distinctly from one another than vowels more centrally located in the perceptual space. As a consequence, French speakers might perceive adjacent point vowels as being more distinct from one another compared with central vowels. This could account for why the found a larger MMF for the French participants compared with the Spanish participants in all the comparisons except in the /ɛ/-/e/ pair, where /ɛ/ is more centrally located, even though /ɛ/-/e/ is a phonemic contrast in French but not in Spanish.

Sharma & Dorman (1999) used consonant-vowel (CV) sequences to better understand the influence of phonetic categories on the MMN. They synthesized nine tokens along a /da/-/ta/ continuum that varied in their duration of VOT (in 10 ms steps, from 0 ms to 80 ms). In a behavioral identification experiment, they found that stimuli with VOT durations between 0 ms and 30 ms were consistently identified as /da/, while stimuli with VOT durations between 50 ms and 80 ms were identified reliably as /ta/. The token with a VOT duration of 40 ms was identified as /ta/ half of the time and /da/ the other half of the time. In the behavioral discrimination task, participants were significantly better at discriminating tokens sampled from opposite sides of the category boundary (30 ms vs. 50 ms), as opposed to tokens sampled from the same side of the category boundary (e.g., 60 ms vs. 80 ms). These results are consistent with the standard findings of the categorical perception of stop consonants based on VOT durations (Liberman, Delattre, & Cooper, 1958; Liberman, Harris, Kinney, & Lane, 1961; see Eimas, 1985 for a discussion of the infant literature on the /t/-/d/ contrast). Measuring electrical evoked potentials, Sharma & Dorman (1999) compared the MMN responses to standard-deviant pairs that either crossed the category boundary (30 ms vs. 50 ms) or were on the same side of the category boundary (60 ms vs. 80 ms). In one block, the 30 ms VOT token was the standard and the 50 ms VOT token was the deviant. In a second block, the 60 ms VOT token was the standard and the 80 ms VOT token was the deviant. The probability of occurrence of the deviant token was 15%. If the MMN was sensitive to phonetic category boundaries, they predicted, then a larger MMN should be elicited for the across-category condition, as opposed to the within-category condition. While they found a minimal MMN in the within category condition, they found a significantly larger MMN in the across-category condition, as predicted. Therefore, while the tokens within each pair were equally distant in acoustic space (20 ms VOT difference between each token in both conditions), the across-category pair elicited a significantly larger MMN than the within-category pair. Therefore, they suggest that they have provided electrophyisiological evidence of the non-monotonic functions found in behavioral categorical perception experiments.

In a subsequent set of experiments, Sharma & Dorman (2000) compared the MMN responses of Hindi and American English speakers on a VOT contrast native to Hindi but not English. They recorded a Hindi speaker producing tokens of /pa/ and /ba/ with pre-voicing[8], similar to that found in Hindi. They manipulated the amount of pre-voicing duration to create a continuum between 0 and -90 ms VOT. English listeners consistently identified the tokens as instances of /ba/, even at the shortest end of the VOT continuum. Hindi participants, however, displayed the canonical non-monotonic categorical perception response function. For the behavioral discrimination experiment and the MMN experiment, they selected the -10 ms and -50 ms VOT tokens. Hindi participants were at ceiling on this contrast, while the English participants were at chance levels of discrimination. In the MMN experiment, the standard was the -10 ms VOT stimulus and the deviant was the -50 ms VOT stimulus, which had a rate of occurrence of 15%. As predicted, a large and reliable MMN beginning roughly 175 ms post-onset of the deviant stimulus was found in the Hindi participants, but absent in the English participants. Consistent with the findings from Näätänen et al. (1997) and Winkler, et al. (1999), Sharma & Dorman (2000) report a reliable MMN to a native language contrast, and a lack of MMN to a non-native language contrast, suggesting that the MMN is sensitive to phonetic category distribu-

[8] It is worth pointing out that the vowels in the experimental stimuli had a slight “r” coloring on the vowel. Moreover, these tokens were easily acceptable to native speakers of Hindi as legitimate lexical items in the language (*baar* “again”; *paar* “side”).

tions of the native language of the participants. An obvious potential confound exists in this experiment. The items presented to participants in this experiment sounded like lexical entries in Hindi (i.e., *baar* and *paar*). Therefore, the possibility remains that the Hindi participants could have simply been mapping the auditory stimulus onto distinct lexical representations, and the fact that they were distinct lexical representations could pose as the source of the MMN. And moreover, while it is difficult to know how closely these items sounded similar to the English lexical items *bar* and *par*, if they did not, then the lack of an MMN in the English participants could be attributed to the fact that they could not be mapped onto (distinct) lexical entries. While this is not likely the case, provided what we know about categorical perception and the response properties of the MMN (see below), this does exist as a possible alternative explanation.

Earlier results from the same group (Sharma, Kraus, McGee, Carrell, & Nicol, 1993) failed to find a clear difference in the modulation of the MMN between the within- and across-category conditions, however. There, they compared the MMN to tokens from the /da/ to /ga/ continuum, manipulating the onset frequencies of the second (F2) and third (F3) formants. In one condition, participants heard standards and deviants of two acoustically distinct tokens of /da/ (the within-category condition), while in a second condition, participants were presented with tokens from across the category boundary (e.g., standard /da/ and deviant /ga/). They failed to find a reliable difference in the magnitude of the MMN when comparing the across-category condition with the within-category condition. They predicted that they should have found a significantly larger MMN to the case where the tokens straddled the phonetic category boundary. On these grounds, they conclude that the MMN is not sensitive to phonetic information, and instead simply encodes acoustic properties of the stimulus (similar conclusions are drawn in Maiste, Wiens, Hunt, Scherg, & Picton, 1995).

While it is clear that to some extent, the studies described above report electrophysiological sensitivity to properties of the stimulus not reflected in their physical attributes (e.g., native language inventory, differential effects contingent upon category boundaries, etc.), it remains to be seen whether or not listeners are constructing phonological representations of the standards in these cases, or whether they are reflecting phonetic category-level distributions. Where the latter to be true, it would not be an insignificant finding. The fact that the MMN/MMF is sensitive to particular properties of the native language phonetic category structure is a noteworthy result. It could have been far less interesting, and the MMN/MMF could simply be responsive to the physical attributes of the signal (equivalent MMNs in within- and across-category boundary contrasts, no effect of vowel inventory size, no difference in the MMN to the phonemic status of a phonetic category in the native language, etc.). Instead, however, it seems that the MMN paradigm is able to elicit sensitivity to these linguistically relevant factors in speech perception, thereby providing a powerful tool in assessing the nature of the constraints and representations at play. In the next series of papers, however, a more linguistically interesting question is asked: can listeners construct representations of the standard in an oddball paradigm at the level of phonology and not phonetics. The inference to be drawn from such a finding would be that auditory cortex supports abstract phonological representations of the input that are independent of phonetic categories, a view not without its opponents (see Johnson, 1997; Pisoni, 1997; Pierrehumbert, 2002; Silverman, 2006; among others).

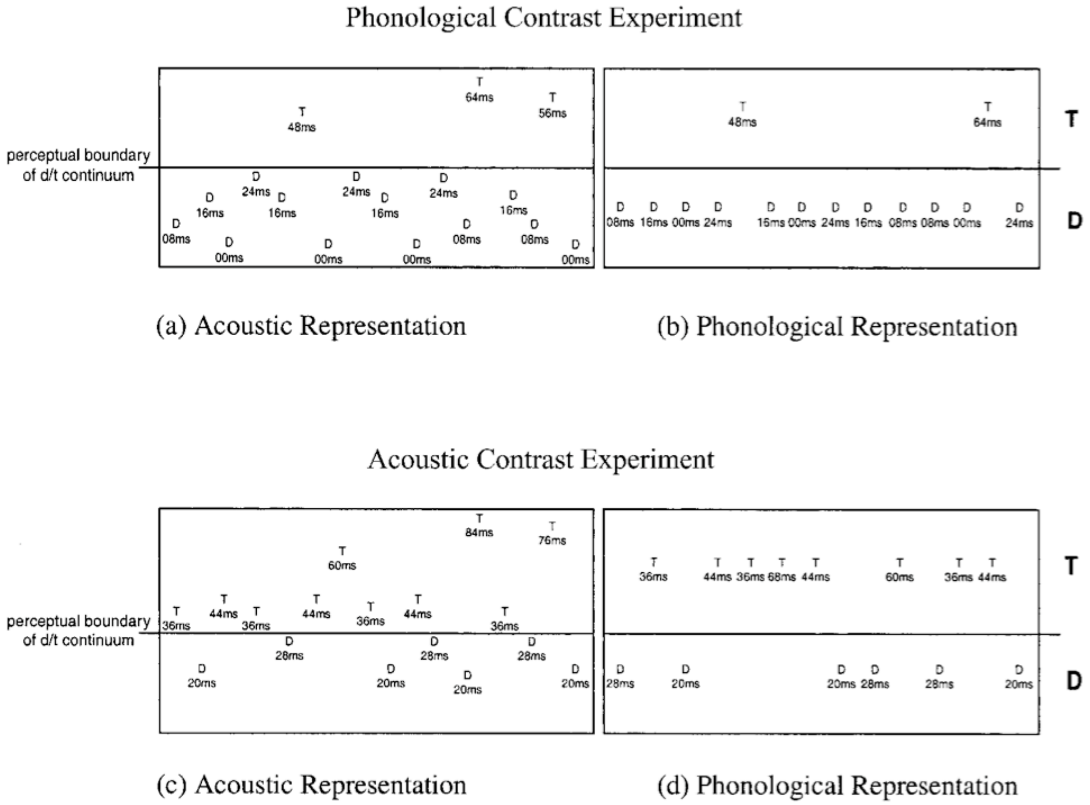
Because the standard was in a one-to-one token-to-category relationship in all of the previously described experiments that looked at speech perception, participants could have simply constructed a memory representation of the standard faithful to its acoustic representation without constructing a higher-level category representation of the standard. In particular, listeners did not need to construct a representation of the standard at the level of phonology. These effects could arise from representations of the standard construct at the phonetic level. The next series of experiments were designed in such a way that the standard varied along some acoustic dimension (similar to the low-level auditory experiments of István Winkler, et al., 1990; Gomes, et al., 1995), forcing participants to perceptually group the standards into a representational object by which the deviant could be compared.

Aulanko et al. (1993) synthesized two-formant syllables that contained either steeply rising or steeply falling F2 formant transitions. These synthetic stimuli contained a short burst of low-frequency energy (35 ms), followed by 50 ms of formant transition and subsequently 35 ms of a steady-state vowel (i.e., /æ/). The F1 transition rose in all tokens from 200 Hz to 700 Hz. The F2 transition, however, was either rising (400 Hz to 1800 Hz) or falling (2800 Hz to 1800 Hz), and it was this difference in the trajectory of the transition that created the difference in perception (a rising transition was perceived as /bæ/ and a falling transition perceived as /gæ/). It should be noted that they did not synthesize a continuum between /b/ and /g/, instead, there were only those two transitions. The crucial manipulation in this experiment was that they synthesized 16 different tokens of these /bæ/ and /gæ/ stimuli, manipulating the fundamental frequency (the perceived pitch): the formants (F1 and F2) and their transitions were held constant, only the pitch of the stimuli were manipulated to create the 16 acoustically distinct tokens. The fundamental frequency (f0) of the 16 variants of /bæ/ and /gæ/ covered a range of over an octave (87 Hz to 204 Hz), with each particular token having a constant f0. Aulanko et al. (1993) claim that the stimuli sounded as though they were being spoken by the same individual. Participants listened passively (they were instructed to read a book of their choice) while neuromagnetic measurements were recorded over the left hemisphere. In one block, participants heard the different pitch variants of /bæ/ as the standard (standards comprised 80% of the presentations within a block) and /gæ/ as the deviant, and vice versa in the other block. Because the different pitch variants were being presented as a standard, an elicitation of an MMF suggests that the listeners are abstracting away from the variance found in the surface acoustic properties of the stimuli and are able to construct a "standard" representation in spite of the variation in pitch by grouping the acoustically distinct tokens of /bæ/ (or /gæ/) together. This is precisely what Aulanko, et al. (1993) find. They report a larger amplitude for the tokens when they are the deviant as opposed to when they are the standard. Therefore, it appears as though the participants in this experiment were able to group together the acoustically different standard tokens to construct an acoustically invariant representation of /bæ/ or /gæ/ (depending on the particular block within the experiment). It is also worth mentioning that they localized both the N100m (M100) and the MMF to supratemporal auditory cortex and did not any reliable differences in source localization between the two components (although they were forced to perform the source analysis with a significantly reduced number of participants from which they calculated the waveform analysis). On the surface, these findings suggest that participants are able to construct an invariant, abstract representation of the repeatedly presented standard independent of the variation in pitch. Given that the spectral properties of the stimuli were identical (for each stimulus type: /bæ/ and /gæ/) above roughly 200 Hz in the Aulanko, et al. (1993) study, Phillips, et al. (2000) caution against drawing too strong of a conclusion that the MMN is, in fact, indexing anything beyond a simple acoustic representation of the standard. If the MMN/MMF is responsive only to spectral information above 200 or 250 Hz, the acoustic variants of /bæ/ are identical, and the variants of /gæ/ are all identical.

To avoid this potential confound, using MEG in an oddball paradigm, Phillips, et al. (2000) also employed a many-to-one design (Gomes, et al., 1995; István Winkler, et al., 1990) in the discrimination of /dæ/ and /tæ/, which differ in the duration of voice onset time (VOT; the duration between the release of the stop closure on /t/ and /d/ and the onset of voicing in the vowel). This time instead of varying pitch (cf., Aulanko, et al., 1993), which is not the primary acoustic/phonetic contrast between /b/ and /g/, Phillips, et al. (2000) modulate the duration of voice onset time (VOT), which is a primary cue in the distinction between /d/ and /t/ (Liberman, et al., 1958). The VOT duration for /d/ is of the range 0 ms to 25 ms and /t/ is of the range 30 ms to 105 ms (Lisker & Abramson, 1964). Liberman, Harris, Kinney, & Lane (1961) demonstrated that for the /t/ versus /d/ contrast, the discrimination abilities of participants is better when the VOTs fall on opposite sides of the category boundaries

(10 ms vs. 30 ms; 20 ms vs. 40 ms) than when they fall on the same side of the category boundary (0 ms vs. 20 ms; 40 ms vs. 60 ms), suggesting that participants are perceiving these acoustically distinct CV tokens in a categorical manner.

As it was mentioned above, Phillips, et al. (2000) employed an oddball design in which there was a many-to-one relationship at the phonological but not acoustic level of representation (cf., Sharma & Dorman, 1999). Not convinced that the previous studies had sufficiently addressed whether the MMN can demonstrate sensitivity to phonological representations, they synthesized a series of stimuli along the /dæ/ - /tæ/ continuum which varied in the duration of the VOT in 8 ms increments. Prior to the MEG recording, each participant was given a forced choice identification behavioral pretest to determine their individual perceptual boundary between /d/ and /t/. After the behavioral pretests, subjects participated in the MEG portion of the experiment, where neuromagnetic signals were recorded while participants listened passively to a series of acoustically distinct tokens. In the first half of the experiment, 87.5% of the acoustic tokens were randomly sampled from one side of the category boundary (12.5% from the other side of the category boundary), and in the second half of the experiment, the majority of acoustic tokens (87.5% again) were randomly sampled from the other side of the category boundary. At the level of acoustic representation, there was no standard, as each successive stimulus presentation was acoustically distinct from the token that either preceded or followed. The "standard" consisted of the range of acoustic tokens from one side of the category boundary, and the deviant was sampled randomly from the other side of the category boundary. Therefore, at the level of category representation, there was a many-to-one relationship, while there was no many-to-one relationship at the acoustic level. The trial structure is provided in Figure 2.8.


Phonological Contrast Experiment

(a) Acoustic Representation

(b) Phonological Representation

Acoustic Contrast Experiment

(c) Acoustic Representation

(d) Phonological Representation

Figure 2.8: Trial structure of Phillips, et al. (2000).

In the phonological contrast experiment, the standards stood in a many-to-one relationship at the level of phonological representation, whereby there were many distinct acoustic tokens as the standards, all of which mapped onto a single hypothesized phonological representation. At the level of acoustic representation, however, there is no standard, as each token is randomly sampled from the set of four tokens from one side of the perceptual boundary. Letters refer to the category to which the token belonged and the number immediately beneath refers to each individual tokens VOT. The trial structures here refer to a schematic representation of the presentation and do not necessarily reflect a particular sequence that occurred in the experiment itself.

In the phonological contrast experiment, Phillips, et al. (2000) report a reliable MMN to the deviant, peaking around 200 ms post-onset of the deviant stimulus, suggesting that listeners were able to construct a category representation at the phonological level for the standard despite the acoustic variation in the individual tokens. That is, listeners seemed to be able to perceptually group these acoustically distinct tokens together to form a category, and that when they perceived a token from the other side of the category boundary, they were able to detect the change (as indexed by the MMN). In order to demonstrate that listeners were actually constructing representations consistent with their linguistic representations and not simply grouping the tokens based on "long" or "short" VOTs, Phillips, et al. (2000) conducted a follow-up experiment whereby 20 ms of VOT were added to all the VOT values, and the same token distribution was presented during the experiment. A schematization is presented in the lower half of Figure 2.8. Consequently, there was no longer a many-to-one relationship at the phonological category level. And as predicted, they did not reliably elicit an MMN, suggesting that in the earlier experiment, listeners were in fact grouping the standards together in a manner consistent with the phonetic space (i.e., the distribution of VOT values relative to the category boundary). The comparison of the MMN elicited in the Phonological Contrast Experiment as opposed to the lack of an MMN in the Acoustic Contrast experiment is presented in Figure 2.9.

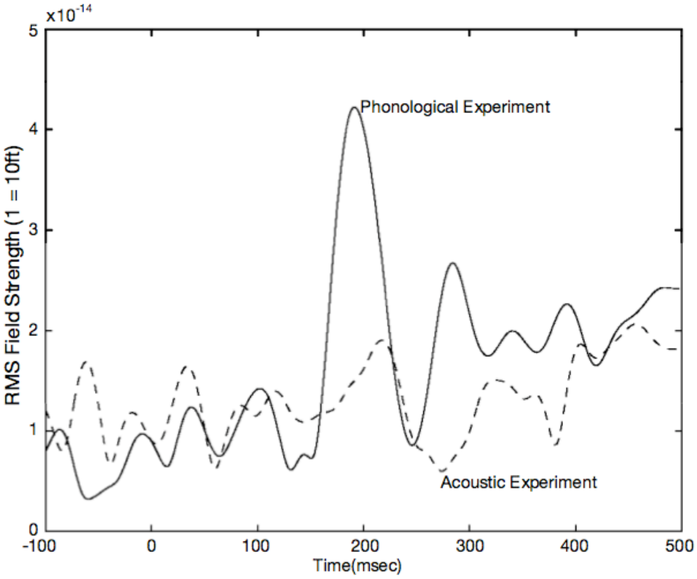


Figure 2.9: Overlay of the difference waves (MMNs) in both experiments.

Solid line refers to the difference wave in the Phonological experiment (standards and deviants separated by the perceptual category boundary), while the dotted line refers to the difference wave in the Acoustic experiment (standards and deviants not separated by the perceptual category boundary).

Ultimately, what this experiment demonstrated is that in an MMN paradigm, participants are able to construct a representation of the standard that is consistent with a category level of representation despite significant variation in the precise acoustic parameter that seems to distinguish the voiced /d/ from the voiceless /t/. And, while it is true that this is quite convincing evidence that the MMN can index some form of abstract representation, these results do not necessarily point to a phonological explanation over a phonetic category explanation. In exemplar models (see Johnson, 1997; Pierrehumbert, 2002), phonetic representations reflect Gaussian distributions along a number of acoustic phonetic parameters. Given that VOT is generally considered to be the primary acoustic cue bifurcating the /t/ - /d/ continuum, we would expect to find a bimodal distribution of VOT values along the VOT continuum (see Figure 2.10) that would serve to underlie the distinction between /t/ and /d/.

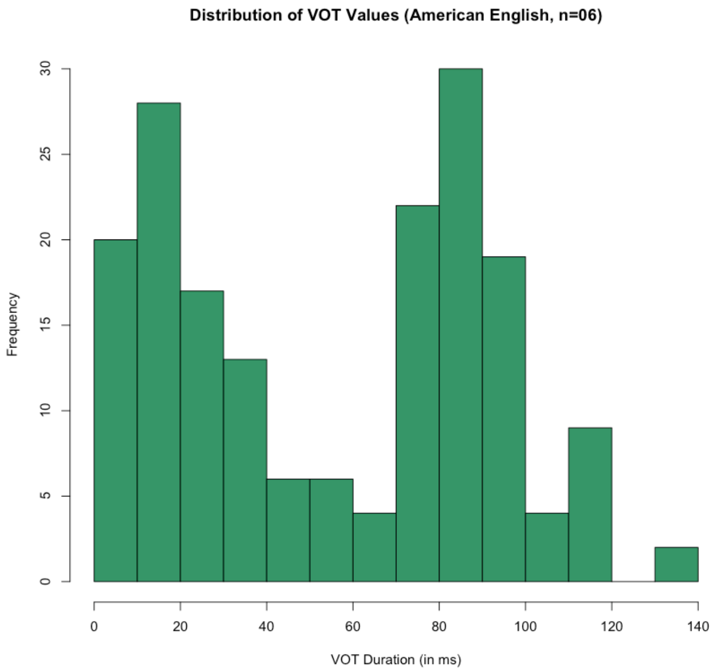


Figure 2.10: Distribution of VOT values for /t/ and /d/

A histogram of distribution of VOT values for /t/ and /d/ in American English. Data were collected from 6 native speakers of American English producing the syllables /dɑt/ and /tɑt/. Notice the bimodal distribution of VOT values (/d/ tokens have a shorter VOT, /t/ tokens having a longer VOT). These data are consistent with previous studies of VOT distributions in English (Lisker & Abramson, 1964; Allen & Miller, 1999).

With this view of the phonetic perceptual space, one could account for the findings in the Phillips, et al. (2000) experiments without invoking a purely phonological level of representation. For example, if participants, when hearing the various acoustically distinct standards sampled from the /d/ side of the perceptual category boundary, were simply mapping all of those tokens into the /d/ side of the VOT distribution, whenever they encountered a VOT that would normally be mapped onto the /t/ side of the VOT distribution, an MMN would be elicited. While this is a rough sketch of an alternative explanation, it serves to point out that while the findings in Phillips, et al. (2000) do point to a conclusion whereby participants are constructing category representations consistent with those involved in linguistic processing, the findings to do not necessarily require a phonological explanation.

Another alternative explanation is one based entirely on neurophysiology. The categorical boundary in English stop consonants is roughly 30ms Voice Onset Time (VOT). The idea is that this VOT boundary is a consequence of auditory neuron responses in quick succession (Sinex & McDonald, 1988; Steinschneider, Schroeder, Arezzo, & Vaughan, 1995). Certain groups of auditory neurons respond to both the noise burst of the stop consonant and the onset of the voicing of the vowel. The refractory period for some of these neurons is roughly 30ms. Thus, it could be the case that the perception of English voiceless stop consonants is a consequence of auditory neurons being able to respond twice because there is enough VOT duration to allow multiple firing. Voiced consonants, however, have a VOT duration less than 30 ms. Auditory neurons are able to fire once. After they fire in response to the burst of the stop consonant, the onset of voicing on the vowel occurs less than 30ms afterward. This occurs during the neurons refractory period, preventing the neuron from responding again. This idea is supported by cellular recordings from non-human primates (Steinschneider, et al., 1995) and other mammals (Sinex & McDonald, 1988).

One of the core definitional properties of being phonological in nature is the direct relationship to meaning (Halle, 2002), however. The phoneme, a unit of representation undeniably phonological, is traditionally defined as the smallest unit of linguistic representation that can serve to distinguish lexical meaning. Therefore, it seems that in order to ensure that we are truly tapping into a phonological level or representation in these types of experiments, it is important to exploit and compare segmental contrasts that do and do not serve to distinguish lexical meanings in different languages. This is precisely what Kazanina, et al. (2006) do.

Again, exploiting the /d/-/t/ VOT continuum, Kazanina, et al. (2006) compare speakers from two languages in a mismatch experiment nearly identical to Phillips, et al. (2000). The real novelty of this experiment is that Kazanina, et al. (2006) tested both Russian and Korean speakers on the /d/-/t/ continuum, speakers of languages which differ in the phonemic status they assign to /t/ and /d/. And moreover, to obviate the aforementioned neurophysiological explanation, Kazanina, et al. (2006) chose languages with shorter VOT boundaries than English. In Russian, both /t/ and /d/ have phonemic status: [tom] 'volume' and [dom] 'house' are two distinct lexical entries, differentiated by the /d/-/t/ contrast. In Korean, however, /t/ and /d/ share an allophonic relationship and appear in complementary distribution. The conditioning environment for this allophonic contrast is as stated: the voiced allophone /d/ occurs intervocalically (e.g., /paTa/ ⟶ [pada] 'sea'), whereas the voiceless unaspirated counterpart occurs word-initially (e.g., /Tarimi/ ⟶ [tarimi] 'iron'). There are two things to note: First, the VOT distributions for Korean and Russian are distinct from English. The perceptual category boundary between /d/ and /t/ in English is roughly +30 ms VOT duration, whereas in Russian, the boundary seems to be in the amount of prevoicing on a particular stop consonant (in the behavioral identification pretest, the boundary seemed to be roughly -16 ms VOT for the Russian participants). The Russian VOT boundary is roughly -15ms VOT. Second, both languages share a bimodal distribution of /d/ and /t/ at the phonetic level. That is, given that both sounds occur in surface representations of speech, they both have some phonetic status in the language. The expectation, then, is that in a many-to-one oddball paradigm using modified tokens of /da/ and /ta/, if participants are constructing a purely phonological representation of the standard, then we expect to find an MMN in the Russian participants but not in the Korean participants, because these sounds are represented independently at the level of phonology in Russian but not in Korean. The distribution of VOT values was -50 ms to +20 ms. In a behavioral discrimination and identification pretest, the Russian speakers showed the typical categorical perception response profile: extremely proficient discrimination ability for tokens from opposite sides of the perceptual boundary and poor discrimination ability for tokens sampled from the same side of the perceptual boundary. The Korean participants, on the other hand, did not as well, judging all the tokens of /ta/ to be a fairly good exemplar of /ta/ (roughly 3 out of 4 on a 0 to 4 scale), and showing no real discernable ability to discriminate between tokens from one side of the category with those from the other side. In the MMF experiment, both Korean and Russian subjects participated. Given that there was a null prediction for the Korean participants, an MMF study using a 1 KHz standard and 1200 Hz deviant was conducted to ensure that an MMF would be elicited in an oddball paradigm in both groups. In the /da/ - /ta/ portion of the MMF experiment, a series of acoustically distinct, yet categorically identical tokens were played as the standard, interrupted by a deviant sampled from the other side of the category boundary. As predicted, Kazanina, et al. (2006) found a reliable MMN for the Russian participants (in the time window of 100 – 180 ms and 180 – 260 ms), and no reliable difference in the RMS of the MEG temporal waveform for the Korean participants in any time window between 20 ms and 340 ms. Unlike the findings from Phillips, et al. (2000), the results in Kazanina, et al. (2006) are considerably more difficult to explain by appealing to phonetic category distributions alone. Given that both [t] and [d] occur phonetically in Korean and Russian, if participants were simply constructing a phonetic representation of the standard, then we would predict that both the Korean and Russian participants to show an MMF to deviant stimuli. However, the lack of an MMF in the Korean participants and the presence of an MMF in the Russian participants suggests that the Russian participants were able to construct a phonological representation of the standard, while the lack of a phonological contrast for the Korean participants prevented them from constructing a phonological representation of the standard. One potential concern interpreting these results stems from the context-dependent distribution of [d] in Korean. The relationship between [t] and [d] in Korean is allophonic in nature and they exist in a complementary distribution: [d] inter-vocalically and [t] everywhere else. If phonetic category distributions are stored with contextual information included regarding allophonic distribution (see Silverman, 2006 for a consistent opinion), then the failure to elicit an MMF in the Korean participants could be explained without referring to phonological structure (and the results for the Russian participants could be explained in a manner similar to the alternative explanation offered for the Phillips, et al. (2000) data). Given that Korean listeners rarely, if ever, hear [d] word-initially (recall that the tokens in the experiment were /da/ and /ta/), their failure to elicit an MMF to the deviant tokens in the experiment could be explained by the fact that [d] never occurs word-initially in Korean, and therefore, word-initial [d] does not exist in the phonetic distribution in the language. This alternative explanation remains a potential way-out for those who wish to maintain a phonetic-centric story for speech perception and the MMN/MMF findings.

Speech sound representations are not, however, 'gestalts'. Instead, a long line of phonological research has demonstrated that phonological segments are further decomposable into distinctive features, abstract articulatory values specifying the production parameters of a given speech sound (Jakobson, et al., 1952; Chomsky & Halle, 1968; Halle, 1983, 1995; Halle & Stevens, 1991; among others). Moreover, given the results from Gomes, et al. (1995), it appears that listeners can perceptually group standards in an oddball MMN design along one particular physical dimension that all the standards share. In a recent experiment, Yeung & Phillips (2004) asked if participants would be able to perceptually group relatively disparate standards along sharing one distinctive feature (i.e., [+voice]). Specifically, does auditory cortex have access to distinctive feature representations? Given that distinctive features are normally stated in articulatory terms, the question of whether they are located in auditory cortex is not a trivial question. Yeung & Phillips (2004) presented participants with a series of CV tokens. The vowel in all trials was /æ/. In 37.5% of the trials, they heard /bæ/, and in another 37.5% of the trials they heard /gæ/. In 12.5% of the trials they heard the pseudo-deviant /dæ/, and in the final 12.5% of the trials they heard the deviant /tæ/. The consonants /b/, /d/ and /g/ all form a natural class: voiced stop consonants. While the consonant /t/ is also a stop, it is produced without vibration of the vocal folds in the glottis. Therefore, in an oddball paradigm, if listeners can construct representations of the standard at the level of the distinctive feature, then they predict to find an MMF to the /tæ/ syllable and not the /dæ/ syllable even though their likelihood of occurrence is identical. Like the Phillips, et al. (2000) and Kazanina, et al. (2006) experiments, Yeung & Phillips (2004) created a many-to-one relationship among the standards. In a behavioral pretest, they gauged the perceptual boundary between the three voiced and voiceless pairs of the 16 participants. Participants were presented with variation in the VOT durations of the standards (four distinct acoustic token per syllable type). By varying the place of articulation of the standards, there was no many-to-one relationship at the phoneme level and no many-to-one relationship at the acoustic level (due to the variation in VOT values). And again, like Phillips, et al. (2000), they increased the VOT values across the board in a second experiment to eliminate the possibility that participants could simply be grouping the standards based on whether they had a long or short VOT. In the MMF experiment, they analyzed two distinct time windows (i.e., early time window: 50 ms – 150 ms; mismatch time window: 150 ms – 250 ms). In the earlier time window, they found a significant interaction between condition (standard/deviant, standard/pseudo-deviant, deviant/pseudo-deviant), region (anterior/posterior) and hemisphere (left/right) only in the acoustic condition and not in the phonological condition. In the mismatch time window, there was a significant interaction for condition × hemisphere in both the phonological and acoustic conditions. Ultimately, they found significant effects in the mismatch region between standards and deviants and also failed to find a difference between the standard and pseudo-deviants in any region. Curiously, however, they also found an effect in the mismatch time window for the acoustic condition (cf., Phillips, et al., 2000). They conclude then, that the MMF is sensitive to the acoustic properties of distinctive features, such as VOT. The incongruence in the results between Phillips, et al. (2000) and Yeung & Phillips (2004) makes these findings difficult to interpret, but highlight quite nicely the power of the MMN/MMF paradigm in trying to assess the representational nature of speech sounds.

One final MMN/MMF experiment I discuss also deals with the nature of phonological feature representations in the lexicon. Eulitz & Lahiri (2004) used the MMN to test whether phonemic representations in the lexicon are underspecified for non-contrastive distinctive feature values in the language. The used the German vowels /e/, /ø/ and /o/ in an oddball paradigm with German participants. The vowel /ø/ in German is a front-rounded mid-vowel. It has been assumed that the feature [coronal] is not specified in the phonological lexicon (Archangeli, 1988; Lahiri & Reetz, 2002). Typically, this claim is made for consonants, and in particular, the nasal consonant /n/, but it has also been extended to front vowels under the assumption that vowels and consonants do not share distinct featural representations for place of articulation. Under this hypothesis, then the vowel /e/ is underspecified for its place of articulation in the phonological lexicon, while /o/ is specified for both [dorsal] and [labial], since it is both round (i.e., [labial]) and back (i.e., [dorsal]). Given that /ø/ is both front and round, it is specified for [labial] but underspecified for [coronal]. Eulitz & Lahiri (2004) also note that the distance between /e/ and /ø/ on the one hand, and /ø/ and /o/ on the other, is roughly equivalent in F2 × F3 formant space. Given that they are all mid-vowels, their value for F1 is roughly equivalent. It should also be noted that

Eulitz & Lahiri (2004) assume (following Lahiri & Reetz, 2002) that [coronal], along with the other features, are specified in the surface, acoustic representation of the vowel tokens. Using an MMN oddball design, they tested the pairs /e/-/ø/ and /ø/-/o/, which each vowel being able to serve as the standard in one block and deviant in the other for each pair. Moreover, they predicted a higher magnitude and earlier peak latency MMN if the mapping involves a feature conflict as opposed to if it involves no conflict (Näätänen & Alho, 1997). The comparison of interest lies in the /o/-/ø/ pair. When /o/ is the standard and /ø/ is the deviant, a conflict at the level of phonological representation occurs. This is because the [coronal] feature extracted from the auditory signal of /ø/ mismatches with the stored representation of [dorsal] for the standard /o/. A contrast should not occur in the opposite direction. If /ø/ is underspecified for its place of articulation, then the constructed representation of the standard does not contain a specification for place, and therefore, the specified [dorsal] feature on /o/ would not conflict. For the /e/-/ø/ pair, since neither is specified for place of articulation, no conflict should exist at the level of phonological representation. Therefore, they predict a larger MMN when /o/ is the standard and /ø/ is the deviant compared to when /ø/ is the standard and /o/ is the deviant. Moreover, they predicted no difference in the /e/-/ø/ pair. They found a clear MMN component in the grand average waveform for all conditions. There was no difference in the latency or amplitude of the MMN for the /e/-/ø/ pair. That is, an approximately equivalent MMN was elicited irrespective of which phoneme was the standard and which phoneme was the deviant. They did, however, find a differential MMN in the /ø/-/o/ pair: a larger and earlier MMN when /o/ was the standard and /ø/ was the deviant than in the opposite configuration. That is, despite the fact that the acoustic difference is identical, a larger and earlier MMN is elicited in one standard/deviant configuration than the other, suggesting that the MMN is indexing more than just the physical properties of the stimulus. Eulitz & Lahiri (2004) suggest that these findings support the predictions of a featurally underspecified lexicon model (FUL; Lahiri & Reetz, 2002; Lahiri, 2007), whereby some features, those that do not play a contrastive role in lexical representation, are not phonologically specified. They reasonably conclude that the asymmetric findings they report are predicted by theories of underspecification.

The point for the current discussion is twofold. First, these results further demonstrate the utility of the MMN/MMF in understanding the nature of linguistic and auditory representations, and additionally, assessing the types of representations to which auditory cortex has access. Second, the findings from Eulitz & Lahiri (2004) further demonstrate that the MMN/MMF can index computations performed over the auditory signal. That is, it does not simply reflect the physical properties of the stimulus (cf., Sharma, et al., 1993; Maiste, et al., 1995). Instead, it seems to be able to index higher-order cognitive representations that, in some cases, are constrained by organizational properties of the native language of the participant (e.g., vowel inventory, phonemic status of a segment, featural representations exploited in phonological inventories, etc.).

For these reasons, the MMN/MMF has proven to be an extremely powerful tool in assessing the types of auditory and linguistic representations supported by auditory cortex. Again, to further reiterate, the MMN/MMF indexes cognitive operations and representations above and beyond the physical properties of the stimulus. In the Sharma & Dorman (1999, 2000) experiments, they found that the magnitude of an elicited MMN is sensitive to phonetic category boundaries of the native language. Specifically, they found that equal acoustic distances (differences in VOT durations) did not equate to equivalent MMN magnitudes. Instead, they found that VOT differences that crossed phonetic category boundaries elicited a larger MMN than inter-category differences. Additionally, the same stimuli elicited different MMNs dependent on the phonetic category status of the native language of the participants. That is, a reliable MMN was found in Hindi speakers for a pre-voicing VOT contrast, while no MMN was found in English participants for the same set of stimuli. A similar conclusion can be drawn from the Näätänen, et al. (1997) and Winkler, et al. (1999) experiments. In these studies, the size/magnitude of the MMN was dependent upon the native language of the participants: the same physical stimulus elicited differential MMN magnitudes dependent upon the native language of the participant. Phillips, et al. (2000) also demonstrated that the elicitation of an MMN was more dependent upon whether the distribution of standards and deviants straddled a category boundary (phonological experiment) than the sheer difference in VOT durations between standards and deviants (acoustic experiment). Kazanina, et al. (2006) also showed differential MMF responses as a function of the native language of the participants to the same physical quantities, while Eulitz & Lahiri (2004) demonstrated that the order of the standards and deviants affected the magnitude and latency of the MMN. The findings from Korzyukov, et al. (2003) demonstrate another (though non-linguistic) case of the MMN indexing constructed representations of the stimulus and not its physical nature. While the ultimate focus of these studies was to investigate the available representations supported by auditory cortex, as well as properties of phonetic category distributions and native language inventories, they also serve to demonstrate that the MMN indexes abstract properties of the stimulus. It should also be noted that the MMN oddball paradigm has been used to investigate the role of phonological constraints on syllable structure in native and non-native speech perception (Dehaene-Lambertz, et al., 2000), the nature of lexical access (Shtyrov & Pulvermüller, 2002; Assadollahi & Pulvermüller, 2003) and certain aspects of syntactic processing (Shtyrov, Pulvermüller, Näätänen, & Ilmoniemi, 2003; Pulvermüller & Shtyrov, 2006).

One limiting condition of the oddball paradigm, however, is that it only seems possible to assess the representational properties of auditory objects. For example, once we demonstrate that abstract phonological representations are supported by auditory cortex beyond a reasonable doubt, it is difficult to imagine how to begin to investigate the set of processes involved in mapping the time-varying auditory waveform onto these abstract linguistic representations. This could be a 'poverty of the imagination' concern. However, given that the core mechanism underlying the MMN/MMF seems to be a comparison metric between an already constructed representation of the standard and the incoming deviant (István Winkler, et al., 1990; István Winkler, et al., 1992; Gomes, et al., 1995; Näätänen, et al., 2007), directly investigating the set of processes that allowed one to arrive at the representation of the standard seems more difficult. Indirectly investigating the set of processes does seem a little more straightforward – if one can determine the relevant properties of the representation objects constructed, then one can more narrowly begin to understand the processes involved in going from the physical acoustic waveform to these higher-order cognitive representations.

In the next section, I review work on an earlier evoked response: the N1/N1m/M100. In the experiments reported in Chapters 3 and 4, I use this evoked response as the dependent measure indexing perceptual processing. The time course of the MMN/MMF is roughly 150 ms – 300 ms post onset of a deviant stimulus. And, by 150 ms – 300 ms, auditory cortex supports phonetic and abstract phonological representations (Sharma & Dorman, 1999, 2000; Phillips, et al., 2000; Eulitz & Lahiri, 2004; Yeung & Phillips, 2004; Kazanina, et al., 2006; Näätänen, et al., 2007). Therefore, if we are interested in beginning to understand the nature of the processes that allow listeners to map the time-varying acoustic waveform onto phonetic and phonological representations, it might be useful to look earlier in the time-course of processing. To date, however, the dependent measures associated with the N1 (and its MEG equivalent: M100/N1m), namely its latency and amplitude, have typically been reported to faithfully reflect physical attributes of the signal, making it appear, on the surface, to be considerably less useful in assessing the types of representations and processes recruited in online auditory and speech perception. In the experiments I report in Chapters 3 and 4, however, it does seem that the M100 is indexing more than just the physical attributes of the signal (though see Roberts, Ferrari, & Poeppel, 1998 for results suggesting some perceptual sensitivity of the M100; these results are discussed further below). Instead, we suggest, it shows sensitivity to hypothesized representations and processes recruited in speech and pitch processing. That is, the

M100 can be an index of the processing of abstract properties of the stimulus relevant to the perception of speech and pitch.

The M100 and N1 in Auditory Processing

In this section, I briefly discuss the M100 (MEG equivalent to the ERP N1), and present a selection of studies demonstrating its sensitivity to physical attributes of the stimulus in both auditory and speech perception. Subsequently, I discuss results that suggest that the M100 can also index perceptual processes. The electrical N1 in EEG is a negative-going potential comprising several subcomponents, with a primary subcomponent localizing to A1 (Picton, Woods, Baribeau-Braun, & Healey, 1976). It is an exogenous response evoked by any auditory stimulus with a clear onset, and is found regardless of the task performed by participant, or his/her attentional state (Näätänen & Picton, 1987). Its MEG counterpart, the N1m or M100, appears to be the magnetic equivalent of the primary subcomponent that localizes to A1 in supratemporal auditory cortex (Hari, Aittoniemi, Järvinen, Katila, & Varpula, 1980; Eulitz, Diesch, Pantev, Hampson, & Elbert, 1995; Virtanen, Ahveninen, Ilmoniemi, Näätänen, & Pekkonen, 1998), thereby making it a more focused dependent measure for use in understanding auditory processing (Roberts, et al., 2000). The M100 is an automatic, auditory evoked neuromagnetic component that peaks roughly 100 ms post-onset of an auditory stimulus (Roberts, et al., 2000). The M100 localizes to planum temporale, a region of auditory cortex just posterior to Heschl’s Gyrus (primary auditory cortex), and with an extremely good signal-to-noise ratio (e.g., 3,600 repetitions per condition), the M100 source localization reflects the tonotopic organization of primary auditory cortex (Lütkenhöner & Steinsträter, 1998). Equivalent Current Dipole (ECD) modeling of the sources of the M100 have also revealed an ampliotopic organization (Pantev, Hoke, Lehnertz, & Lütkenhöner, 1989; cf., Vasama, Mäkelä, Tissari, & Hämäläinen, 1995).

The primary dependent measures of the M100 that have been exploited are its peak latency (sometimes the peak latency of the RMS), the duration of time between stimulus onset and the point of maximal magnetic field strength elicited by a sound, and peak amplitude, the maximum magnetic field strength at the peak latency. The typical duration between stimulus onset and peak magnetic field strength is 100 ms, though a range of values between 80 ms and 150 ms are not uncommon, and we find significant variation in the raw peak amplitude values across different participants.

Evidence suggests that the M100 integrates over the first 25 ms to 45 ms of stimulus onset. Forss, Mäkelä, McEvoy, & Hari (1993) presented participants with a series of click trains at four different rates (40, 80, 160, 320 Hz) with a duration of 200 ms and ISIs of 1 sec and 4 sec. They found that the peak latency decrease of the M100 was commensurate with the shortening of the click intervals down to as low as 40 Hz. Forss, et al. (1993) understood these findings to suggest that the temporal integration window for the M100 is roughly the first 20 to 25 ms of stimulus onset. Gage & Roberts (2000) presented participants with sinusoids that varied in duration, and found that the amplitude of the M100 scaled with stimulus durations up to 40 ms, at which point it saturated, and failed to increase as duration increased. These results suggest that the M100 is able to integrate sensory information over the first 40 ms of auditory stimulus. Finally, Gage, Roberts, & Hickok (2006) presented participants with 1 KHz sinusoids with silent gaps inserted at either 10 ms post-onset of the 1 KHz sinusoid or 40 ms post-onset of the sinusoid. The gaps varied in duration from 0 ms to 20 ms (0, 2, 5, 7, 10, 15, 20 ms). They predicted that, if the M100 does, in fact, integrate only over the first 40 ms of stimulus onset, then the response properties (latency, amplitude) of the M100 should by affected by gaps inserted 10 ms post-onset, but not by gaps inserted 40 ms post-onset, which is beyond the hypothesized temporal integration window of the M100. In the 10 ms condition, both the latency and amplitude scaled linearly dependent on the duration of the gap: M100 latency increased as gap duration increased and M100 amplitude decreased as gap duration increased. It is worth noting, that gap durations as short as 2 ms had detectable impacts on the response properties of the M100, suggesting that the neurobiological generators of the M100 are sensitive to even the shortest discontinuities in the signal, at least within its temporal window of integration. In the 40 ms condition, however, gap duration had no discernable effect on the latency or the amplitude of the M100. Collectively, this set of results suggest that the temporal integration window of the M100 is the first 25 ms to 40 ms of stimulus onset, and that modulations of the spectral properties of an auditory stimulus outside of this temporal window of integration cannot be indexed by the response properties of the M100.

A primary physical attribute modulating the latency of the N1 and M100 is stimulus frequency, or pitch. Jacobson, Lombardi, Gibbens, Ahmad, & Newman (1992) demonstrated frequency specific effects on the N1e (ERP N1; MEG M100/N1m). Participants were presented with 250 Hz, 1000 Hz and 4000 Hz sinusoids. They report a longer N1e latency to the 250 Hz sinusoid compared with the 1

KHz and 4 KHz sinusoids. Using MEG, Roberts & Poeppel (1996) presented five participants with eight distinct pure sinusoids, each of which was 400 ms in duration, and the sinusoids had frequencies of 100, 200, 500, 1000, 2000, 3000, 4000, 5000 Hz. Participants listened passively while neuromagnetic potentials were recorded over the scalp. The source of the M100 localized to superior temporal cortex, although they did not have a sufficient signal to noise ratio to reliably detect a tonotopic organization of ECD sources. They did, however, find that the latency of the M100 varied as a function of stimulus frequency. The mid-range stimuli (500 Hz, 1000 Hz, 2000 Hz) had the shortest M100 response latencies, while the low- and high-range stimuli evoked successively longer M100 latencies in a “quasi-parabolic function”. Though not discussed in Roberts & Poeppel (1996), the amount of variance in the latency of the M100 across participants is reduced in the responses to stimuli closer to 1000 Hz. The M100 response latencies in Roberts & Poeppel (1996) are presented in Figure 2.11.

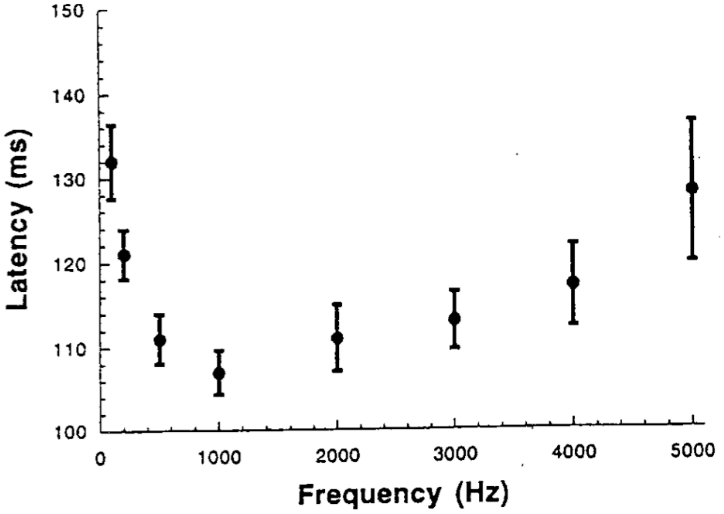


Figure 2.11: M100 response latency as a function of stimulus frequency

The latency of the auditory evoked neuromagnetic M100 as a function of stimulus frequency. Stimuli with frequencies closest to 1000 Hz appear to elicit the shortest latency. Reprinted from Roberts & Poeppel (1996).

Roberts, et al. (2000) report unpublished data that also demonstrates the relationship between stimulus frequency and M100 latencies. They presented pure sinusoids with frequencies of 100, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750 and 1000 Hz. Again, they replicated the finding that lower frequency tones (< 500 Hz) elicited longer M100 latencies than higher frequency sinusoids, but also noted that the latency of the M100 did not differ for sinusoids between 500 and 1000 Hz. They reliably fit the response latency of the M100 to different frequencies with an asymptotic curve (A + B/f; A is the asymptotic latency; B is a scaling constant; f is the frequency of the sinusoid).

Stimulus intensity has also been shown to have an impact on both the amplitude and latency of the M100. Elberling, Bak, Kofoed, Lebech, & Saermark (1981) presented participants with 1 KHz and 500 Hz sinusoids at a variety of different intensity levels: 5 to 85 dB HL. They found decreasing the stimulus intensity led to a delayed latency and attenuated amplitude in the evoked neuromagnetic response. Vasama, et al. (1995) presented six participants with 200 ms 1 KHz sinusoids at four different intensity levels (40, 50, 60, 65 dB HL). They found that the M100 latency decreased and amplitude increased as a function of stimulus intensity (shorter latencies and larger amplitudes for stimuli with greater intensities). Stufflebeam, Poeppel, Rowley, & Roberts (1998) presented participants with a series of different sinusoidal frequencies (100, 200, 2000 and 3000 Hz) at different intensity levels (0, 5, 10 and 20 dB sensation level (SL)). They found that the M100 response amplitude reliably increased with a corresponding increase in stimulus presentation intensity and that the M100 response latency reliably decreased when stimulus presentation intensity increased. Moreover, across the three different peri-threshold presentation levels (5, 10, 20 dB SL), they replicated the M100 response latency curve based on the frequency of the stimulus (e.g., Elberling, et al., 1981; Roberts & Poeppel, 1996).

Given that we rarely perceive pure sinusoids in our natural environment, understanding how auditory cortex, as indexed by these early evoked responses, analyzes more complex sounds might prove useful in determining the relevant features of the auditory signal and the perceptual properties auditory cortex exploits in processing. Roberts, et al. (2000) present additional unpublished data demonstrating the latency of the evoked M100 response is sensitive to more sophisticated properties of auditory objects than simply their frequency and perceived intensity. In one experiment, they presented participants with 100 Hz and1 KHz sinusoids and found a delayed 30 ms latency difference to the 100 Hz sinusoid compared with the 1 KHz sinusoid. They presented the same participants with amplitude-modulated tones with a carrier 1 KHz sinusoid with a modulation frequency of 100 Hz. With a modulation depth of 200%, the frequency spectrum consisted three components (900, 1000 and 1100 Hz) with equal energy. They report that there was a clear contribution of the 100 Hz in the subjective perception of the stimulus, despite the fact that there was no acoustic energy at the modulation frequency (100 Hz). The M100 latency to the AM tone was significantly longer than the latency to the 1 KHz pure sinusoid, but also significantly shorter than the M100 latency to the 100 Hz pure sinusoid. In another unpublished experiment, Roberts, et al. (2000) report that the fine structure of auditory stimulus also modulates the M100. They presented participants again with 100 Hz and 1 KHz sinusoids, but also presented them 100 Hz triangle and square waves. Crucially, they note, while the pure sinusoids have harmonics only at the center frequency of the stimulus, triangular and square waves contain higher harmonic components. Triangle waves have frequency components in all harmonics (H1, H2, H3, H4, etc.), while square waves have frequencies present only at the odd harmonics (H3, H5, H7, etc.). They report a reliably shorter M100 latency to the 100 Hz square wave than to the 100 Hz triangle wave. If the M100 was sensitive to the spectral density of the tokens, one might expect no difference between the two, since they had approximately equal energy distributions, since the harmonic components are evenly distanced across the same range in the frequency domain. They attribute the difference in M100 latency to the fact that the triangle wave has a lower frequency periodicity (since all the harmonics are present, and not only the odd harmonics) compared to the square wave. Given that lower frequencies elicit shorter M100 latencies (e.g., Elberling, et al., 1981; Roberts & Poeppel, 1996), the triangle wave is predicted to elicit longer M100 latencies than the square wave.

The N1/M100 has also been used in the study of lower level acoustic-phonetic properties of speech. Some of these studies are reviewed in Chapters 3 and 4, but it useful to discuss them now as well. In vowel perception, the M100 has been shown to reliably index the value of F1. First formant frequencies closest to 1 KHz elicit the shortest M100 response latencies. Diesch, Eulitz, Hampson, & Ross (1996) synthesized four German vowels (/a/, /i/, /u/, /æ/). The only difference they found in the latency of the evoked M100 was that the low vowels (/a/ and /æ/) elicited reliably shorter M100 latencies than the high vowels (/i/ and /u/). The F1 for the low vowels is considerably closer to 1 KHz (F1; /a/: 780 Hz; /æ/: 600 Hz) than the F1 for the high vowels (F1; /i/: 250 Hz; /u/: 250 Hz). Given that they did not find a difference between the front and back vowels (front vowels have a higher F2, back vowels have a lower F2), a reasonable attribution to the difference in M100 latencies is that the M100 is preferentially responding to F1. Poeppel, et al. (1997) created tokens of three synthetic American English vowels (/i/, /u/, /a/) and found a reliably shorter M100 latency to /a/ than /u/, and did not report a difference between /u/ and /i/. Govindarajan, Phillips, Poeppel, Roberts, & Marantz (1998) replicated these findings again using English participants, and also extended the results to show that a similar patterns of effects are found for non-speech sinusoidal tone complexes. Tiitinen, et al. (2005) replicated these findings in Finnish speakers, showing again that /a/ elicits faster M100 latencies than /u/ using semi-synthetic speech. In the perception of CV syllables, it has been shown that stop consonants (e.g., /b/, /d/, /g/, /p/, /t/, /k/) elicit shorter M100 latencies and faster reaction times than non-stop consonants (e.g., fricatives: /f/, /s/; sonorants: /m/, /l/, /r/, etc.). Gage, Poeppel, Roberts, & Hickok (1998) presented participants with forty single-syllable words that were matched on form class, voicing and rhyme and varied only in the initial consonant of the stimulus. Half of the words began with one of the stop consonants listed above, while the other half began with one of the non-stop consonants listed above. They found a reliable effect of the consonant type on the M100 latency: stops elicited shorter M100 latencies than non-stops. They also found a consonant-type by hemisphere interaction: stop consonants elicited shorter latencies in the right hemisphere than in the left, and non-stop consonants elicited shorter latencies in the left hemisphere than in the right. Moreover, they found a main effect of consonant-type on the amplitude of the M100: stop consonants elicited larger M100 amplitudes in both hemispheres than non-stop consonants. Gage, et al. (1998) attribute the pattern of results they find to differences in the onset dynamics of the two classes of consonants. Stop consonants typically contain greater overall energy and a faster transition to peak energy than non-stop consonants.

Finally, VOT durations have also been shown to impact N1 and M100 response properties. In the MMN studies of Sharma & Dorman (1999, 2000), they also report the impact of VOT durations on the evoked latencies of the N1 (and N1'). Recall that in the MMN portion of Sharma & Dorman (1999), participants were presented with two pairs of syllables: 30 ms – 50 ms VOT and 60 ms – 80 ms VOT on coronal stops. They found a larger MMN to the across category oddball condition (30 ms – 50 ms VOT) than in the within-category oddball condition (60 ms – 80 ms

VOT). In addition to the MMN experiment, they also presented participants with 300 tokens each of the nine stimuli from the VOT continuum they synthesized. They measured the latency of the evoked electrical N1 component. For the tokens with a short VOT (0 ms – 40 ms VOT), they report a single negativity peaking around 100 ms post-onset of the target (N1). Interestingly, however, for the syllables with longer VOTs (≥ 50 ms VOT), they report two negative deflections in the grand averaged ERP waveforms: an earlier negativity (N1') and another negativity peaking approximately 200 ms post-onset of the stimulus (N1). The grand averaged waveforms from Sharma & Dorman (1999) illustrating the presence of the N1 in the short VOT tokens and the additional presence of the N1' in the longer VOT tokens are presented in Figure 2.12.

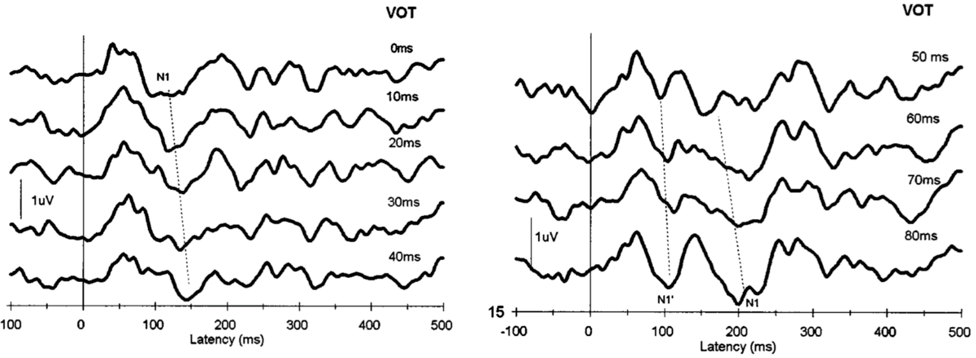


Figure 2.12: Grand average ERP waveforms showing N1 and N1'
Grand averaged ERP waveforms showing the response in the N1 experiment to /ta/-/da/ tokens that varied in their duration of VOT. Notice the single negativity in the shorter VOT conditions, and the presence of an additional negativity in the longer VOT conditions. Reprinted from Sharma & Dorman (1999).

Also relevant in their findings is that they found a high correlation (*r* = 0.84) between VOT duration and peak latency of the N1 component. They failed, however, to find a similarly strong correlation (*r* = 0.03) in the peak latency of the N1' component, suggesting that perhaps the N1’ is a simple index of the stop burst at stimulus onset. They also note that the split between tokens with an additional negativity compared to those without parallels the perceptual split between /da/ and /ta/. Stimuli generally perceived as /da/ elicited one negativity, while stimuli perceived as /ta/ elicited two negative components. Ultimately they speculate that this pattern of results is consistent with the N1 indexing sensory encoding in the neuronal population level (see Sinex & McDonald, 1988). In Sharma & Dorman (2000), they present findings from a similar experiment, except this time they tested Hindi and English speakers on a Hindi VOT contrast (0 ms – -90 ms VOT). Along with the MMN experiment, which showed that Hindi participants elicited a reliable MMN on the Hindi contrast while English participants did not, they also presented both groups of participants with 300 repetitions of each of the 10 tokens synthesized along the pre-voicing VOT continuum. In the N1 analysis, they found only one negative going component (N1), which peaked roughly between 150 ms and 225 ms post-onset of the stimulus. In both groups of participants they again found a strong correlation of VOT duration and N1 evoked latency (Hindi: $r = 0.79$; English: $r = 0.8$). Frye, et al. (2007) used MEG to determine whether the latency of the M100 is also modulated reliably by VOT duration. They presented eight native speakers of American English with tokens from an 11-step VOT continuum between /ba/ and /pa/. Based on findings from a previous set of behavioral experiments with the same stimuli (Liederman, Frye, Fisher, Greenwood, & Alexander, 2005), the tokens from along the 11-step continuum could be categorized into four groups: /ba/ (0 ms – 10 ms), boundary (15 ms – 20 ms), short /pa/ (25 ms – 30 ms) and long /pa/ (40 ms – 50 ms). The calculated the ECD waveform (strength of Equivalent Current Dipole over time), and found a reliable difference in the latency and amplitude of the M100 dependent upon the VOT duration: shorter VOTs elicited faster M100 latencies and greater M100 amplitudes in the left hemisphere only.

Provided the results above, it seems clear that the N1/M100 has typically been attested to index primary and secondary acoustic features of the auditory stimulus. In a handful of cases, however, it has also been suggested to index more perceptually driven processes. For example, the M100 has been shown to index a missing fundamental component (Pantev, Hoke, Lütkenhöner, & Lehnertz, 1989; Fujioka, et al., 2003; Monahan, de Souza, & Idsardi, 2008). Under certain circumstances, listeners can fill in the fundamental component of a complex sinusoid if it is missing (cases of "missing fundamental", "virtual pitch", etc.). The most common ecologically relevant example of this is conversation on the telephone: band-pass filtering imposed by mechanical constraints on the conductance of speech over telephone lines eliminates frequencies below 400 Hz and above 4 KHz. Listeners, however, have no subjective difficulty in reconstructing the pitch of the talker despite its absence in the physical signal. In short, in the series of experiments listed just above, the M100 seems to be more sensitive to the perceived virtual pitch than the physical properties of the stimulus. Further discussion of this topic is presented in Chapter 3 (published as Monahan, et al., 2008), along with a criticism of Pantev, Hoke, Lütkenhöner, et al. (1989) and Fujioka, et al. (2003), which suggests that their modulation of the M100 could be the result of potential confounds, and that it could be driven by physical properties of the stimuli that were used in those experiments.

Roberts, et al. (1998) presented participants with two-tone stimuli that were mixtures of two sinusoids: 100 Hz and 1 KHz. The individual sinusoids varied in their relative amplitude between pure 100 Hz (presence of 1 KHz was not detectable) and pure 1 KHz (presence of 100 Hz was not detectable). In a psychophysical pretest, participants were given a two alternative forced choice task and asked to respond whether they perceived the stimulus as being predominately low frequency or predominately high frequency. Interestingly, they found a non-monotonic response function: stimuli with a -60 dB difference (rate of low frequency categorization subtracted from rate of high frequency categorization; 100 Hz – 1 KHz) to -6 dB difference were all categorized as being predominately high frequency, while stimuli with a +6 dB difference to +80 dB difference were all equally categorized as being predominately low frequency. In between -6 dB and +6 dB, there was a crossover in the response from being predominately high to predominately low. This is presented in Figure 2.13.

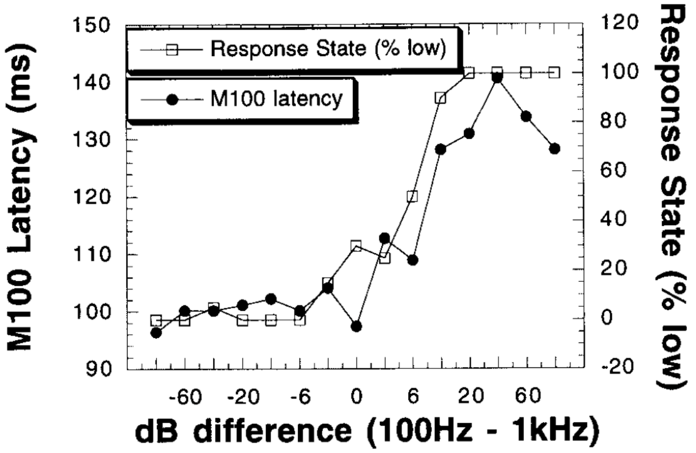


Figure 2.13: Overlay of Psychophysical responses and MEG latencies to tone-mixtures in Roberts, et al. (1998).

Comparison of the psychophysical responses (open squares; single representative subject) and M100 latencies (closed circles) to the 100 Hz – 1 KHz tone complexes that varied in the relative intensity of each frequency component. Notice the strong non-monotonic response function in the psychophysical data. The M100 latencies approximately pattern with the psychophysical responses. Reprinted from Roberts, et al. (2000).

The latency of the M100 in Roberts, et al. (1998), while not as smooth as the psychophysical response, did illustrate a general ‘categorical perception’ curve (as shown in Figure 2.13), suggesting the response latencies of the M100 in this particular experiment are better modeled by the psychophysical responses than properties of the physical stimulus. One final example of a case where the M100 response seems to better track perception than physical attributes is from a study on vowel perception. Roberts, et al. (2004) showed the latency of the M100 to the first formant (F1) respects vowel category boundaries. They created tokens of /a/ and /u/ and modulated F1 in 50 Hz increments between 250 Hz and 750 Hz. The M100 latencies clustered into three distinct bins: the lowest F1 values (250 – 350 Hz) elicited the longest latencies, the middle F1 values (400 – 600 Hz) elicited reliably shorter latencies and the high F1 values (650 – 750 Hz) elicited even shorter M100 latencies. The lowest bin (250 – 350 Hz) generally represents the range of F1 values found for /u/ in natural productions of English, while the highest bin (650 – 750 Hz) generally represents the range of F1 values found in natural productions of /a/. Given that the M100 does not follow a linear response function in this experiment, a plausible conclusion to be drawn is that the M100 is also sensitive to the acoustic distributions of phonetic vowel categories. This experiment is discussed in more detail and a similar conclusion is drawn from the findings presented in Chapter 4.

In short, the N1/M100 reliably encodes both primary and secondary physical properties of the stimulus. A primary property of the auditory stimulus, such as frequency and intensity, modulates the latency and amplitude of the M100. Secondary properties, such as modulating frequencies and fine structure of the auditory waveform, also modulate the latency and amplitude of the M100 in a predicted manner. In experiments on speech perception, the first formant (F1) in vowel perception, the acoustic characteristics of the onset of consonants (stop vs. non-stop), and the duration of VOT all systematically modulate the latency of the N1/M100, suggesting that it is particularly adept at indexing physical properties of the auditory stimulus. Finally, a handful of experiments have shown some level of perceptual influence on the response properties of the N1/M100. The experiments that follow in Chapters 3 and 4 provide further evidence that the M100 is, in fact, modulated by perceptual cognitive processes in a series of experiments where the physical properties of the experimental stimuli have been controlled to eliminate the possibility that the pattern of effects found are due to a confound in the stimulus construction (cf., Fujioka, et al., 2003; Pantev, Hoke, Lütkenhöner, et al., 1989).

Discussion

It has been a hallmark of phonological theory to posit that the representational units of computation are discretized, abstract segments and distinctive features (Jakobson, et al., 1952; Chomsky & Halle, 1968; K. N. Stevens & Halle, 1967; Halle, 1983; Halle & Stevens, 1991; Halle, 1995). Abstract segments (or phonemes) are, in some sense, epiphenomenal: phonological representations consist of a hierarchically structured set of distinctive features bound at the top by one unique node. These distinctive features provide articulatory commands ([+voice] signals to the articulation system that the vocal folds need to be vibrating during the production of this phonological segment) and can be the target of phonological operations (McCarthy, 1988; Halle, 1995). Compelling evidence for these phonemes and distinctive features from psychology and neuroscience, however, has not been easy to come by (Sussman, 2000).

Classical models of speech perception sought to understand the processes and representations that sub-serve the online mapping from a time-varying acoustic waveform onto long-term stored memory representations (Joos, 1948; Halle & Stevens, 1962; Liberman, et al., 1967). Models have varied in their assumptions regarding the nature of the stored representations, whether they be fundamentally auditory (Greenberg, 1999, 2006; Guenther, 2002) or articulatory (Liberman & Mattingly, 1985; Fowler, 1986). Moreover, the nature of lexical representations has been a hotly debated issue. A fundamental cut in the different theories is between those that believe there are abstract and discrete lexical representations with those that believe lexical representations are simply amalgamations of tokens stored in a multi-dimensional space (Goldinger, 1996b; Johnson, 1997; Silverman, 2006). The importance of this distinction for the present dissertation is that the nature of lexical representations constrains the types of representations and processes implicated in earlier, pre-lexical stages of processing.

As it was discussed at the outset of this Chapter, there is ample evidence that abstract representations are computed in the course of speech perception (see Obleser & Eisner, 2009 for a review of recent neurophysiological evidence in favor of this perspective). Subsequently, we presented two different electrophysiological components: the later MMN/MMF and the earlier N1/M100. The advantage of using electrophysiology over behavioral measures is that it allows us to tap into the earliest stages of auditory processing in a completely non-invasive manner. Moreover, results typically reflect the output of extremely early, automatic perceptual processes that are not clouded by particular task demands or controlled processes. A variety of MMN/MMF studies showed listeners ability to construct invariant representations over variable acoustic input (István Winkler, et al., 1990; Gomes, et al., 1995) or infer relational rules about a sequence of independent tones (Korzyukov, et al., 2003). Moreover, I reviewed a series of MMN/MMF experiments that suggest that auditory cortex is sensitive to native language phonetic category distributions (Sharma & Dorman, 1999, 2000; Alho, Sainio, Sajaniemi, Reinikainen, & Näätänen, 1990; Näätänen, et al., 1997; István Winkler, et al., 1999), and more interestingly, can support representations at the level of phonemes (Phillips, et al., 2000; Kazanina, et al., 2006) and even perhaps distinctive features (Yeung & Phillips, 2004). While the MMN is an exceptionally powerful tool for assessing the types of sensory objects participants can treat as equivalent, and thereby infer the types of representations supported by auditory cortex (in auditory MMN/MMF experiments), understanding the processes that allow listeners to arrive at the more complex representations remains more difficult with this measure. Therefore, I suggest to look earlier in the evoked electrophysiological signal to be able to begin to assess early representations and processes that may be involved in mapping the acoustic waveform onto long-term memory representations. Subsequently, I reviewed a series of N1 and M100 experiments that demonstrate that component's sensitivity to physical attributes of the stimulus (frequency, intensity, amplitude modulation, fine-structure of the waveform, F1, VOT, etc.). While evidence of early auditory abstraction is scant in the N1/M100 literature, there are some findings that point in this direction (Roberts, et al., 1998; Roberts, et al., 2004).

In the next three Chapters, I present a series of MEG experiments that demonstrate early auditory abstraction in pitch processing (Chapter 3), vowel perception and normalization (Chapter 4) by exploiting the latency of the M100, and while not exploiting the M100, evidence that knowledge of the phonological structure of the language constraints early (~150 ms post-onset) parsing of the speech signal (Chapter 5).

Chapter 3: Early Auditory Restoration of Fundamental Pitch[*]

Introduction

Pitch is the perceptual correlate of the fundamental periodic component of an auditory signal (F_0). An accurate encoding of the information carried in the fundamental component is required for the successful perception of various kinds of linguistic and paralinguistic information (e.g., lexical tone, intonation, voicing, and speaker identification and emotional state) and non-linguistic auditory input (e.g., music perception). Listeners are adept, however, at recovering the fundamental component from alternative regions of frequency space when the fundamental component itself is missing or masked (Schouten, Ritsma, & Cordozo, 1962; Schouten, 1970; Terhardt, 1974; Smoorenburg, 1970). One everyday example of this effect can be observed with adult voices transmitted telephonically: the fundamental component of the voice is typically below 300 Hz, but narrowband digital telephony transmits only between 300 - 3400 Hz. Consequently, the listener must *reconstruct* the pitch from the signal in the passband. Given the relative importance of its contribution, recovering the pitch of a signal is integral for constructing a holistic percept for a given auditory stimulus and ultimately arriving at the recognition of an auditory object. The present study uses magnetoencephalography (MEG) to measure an early, automatic

[*] Published as:

Monahan, Philip J., Kevin de Souza and William J. Idsardi (2008). Neuromagnetic Evidence for Early Auditory Restoration of Fundamental Pitch. *PLoS One* 3(8): e2900.

evoked auditory response, the M100 (or N1m), building on and extending some previous studies that required some clarification. I find that the M100 latencies of the inferred pitch stimuli match those evoked by actual sinusoidal tones with the same frequencies, suggesting that inferred pitch is recovered by 100 ms and, moreover, that the M100 encodes computations performed over the input and not just transparent spectral properties of the stimulus.

The neural mechanisms that reconstruct the lower end of the frequency spectrum and reconstitute information present in the fundamental component are still largely unknown (see Shamma & Klein, 2000 for models; Goldstein, 1973). Listeners' ability to reconstruct this spectral information, and in particular, to recover the fundamental component (F_0), has been termed fundamental restoration, also known as *inferred pitch*, the *missing fundamental* phenomenon or *virtual pitch* (Goldstein, 1973). This phenomenon has also been observed in non-human mammals (Bendor & Wang, 2005; Cedolin & Delgutte, 2005; Heffner & Whitfield, 1976; Tomlinson & Schwartz, 1988). From a neurophysiological perspective, understanding the time course of fundamental restoration is a prerequisite to identifying the range of neurobiological mechanisms potentially responsible for the reconstruction of the fundamental component.

Recently, the temporal and spatial dynamics of fundamental restoration have been explored using electrophysiology (Fujioka, et al., 2003; Pantev, Hoke, Lütkenhöner, et al., 1989; Matsuwaki, et al., 2004; I. Winkler, Tervaniemi, & Näätänen, 1997). The focus of this work has been on determining the neuroanatomical basis of fundamental restoration. In particular, by employing source-localization analysis of the M100, the fundamental restoration has been localized to the transverse temporal gyrus and the superior temporal gyrus (Matsuwaki, et al., 2004).[9] Moreover, independent neural generators appear to underlie the perception of pure sinusoids and their inferred fundamental counterparts (Fujioka, et al., 2003). In an attempt to understand the temporal dynamics of fundamental restoration, Winkler and colleagues found no latency or amplitude differences using EEG in the N1 between spectral and restored fundamental stimuli (I. Winkler, et al., 1997). The only differences they found were to tokens with long durations (500 ms, as opposed to 150 ms in duration) in a mismatch negativity paradigm.

Perhaps most notably, Pantev and colleagues used MEG and compared the neuromagnetic responses to two sinusoids (250 Hz and 1000 Hz) and a tone complex with an inferred pitch of 250 Hz (1000 Hz, 1250 Hz, 1500 Hz and 1750 Hz) (Pantev, Hoke, Lütkenhöner, et al., 1989). Presenting a source-based analysis of the MEG responses, they concluded that the neural generators of the M100 reflect the processing of the subjective perception of the pitch of a stimulus and not the actual stimulus properties. In other words, the neuronal computations required to reconstruct the fundamental component are performed within 100 ms post onset of the target and reside in early auditory cortex. While the evidence I present here is consistent with this conclusion, there are some caveats that should be noted regarding their findings. First, for

[9] This is the reasoning employed by Matsuwaki, et al. (2004). The possibility exists, however, that the computations subserving the extraction of the missing fundamental component operate at earlier stages of auditory processing and that the output of these computations are simply reflected in the M100 (consistent with physiological models of missing fundamental computation that propose the harmonic templates responsible for inferential pitch extraction occur proir to cortical processing, see Shamma & Klein, 2000 for a similar such model).

the tone complex used in their study, they inserted a continuous band-pass noise centered at 250 Hz, essentially building an equivalent actual pitch into the stimulus that was intended to elicit an inferred pitch. The findings would have been much more convincing had they used a broader band of noise as a spectral masker, say from DC to 500 Hz. Second, the sampling rate for the early MEG equipment was coarse (250 Hz), thereby making it difficult to assign an interpretation to the latency data. The reported latency differences were 4 ms, or one sample at this sampling rate.

Independent research on the M100 suggests that its latency is modulated by spectral characteristics of auditory input. In particular, M100 response latencies are shortest to sinusoids with a frequency of 1000 Hz and longer to frequencies both above and below 1000 Hz (i.e., forming a parabola centered near 1000 Hz) (Roberts & Poeppel, 1996). Therefore, if the neuromagnetic signal was, indeed, primarily reflecting the reconstruction of a fundamental component, then I should expect the latencies for the 250 Hz sinusoid and the tone complex with a 250 Hz inferred pitch to have roughly the same latency, and both should be significantly longer than the M100 response to the 1000 Hz sinusoid. This straightforward prediction is only borne out in two of the six participants reported in the Pantev study (Pantev, Hoke, Lütkenhöner, et al., 1989). In a more recent electrophysiological study investigating the neurobiological properties of fundamental restoration, Fujioka and colleagues (Fujioka, et al., 2003) compared neuromagnetic responses to tone complexes with inferred fundamentals of 250 Hz, 500 Hz and 1000 Hz composed of their 2nd through 5th harmonics, 6th through 9th harmonics and 10th through 13th harmonics. They report that all stimulus parameters (periodicity, harmonic order level, stimulus type (pure tone, inferred fundamental inducing tone complex)) affected M100 latency.

It is also known that the M100 response latency is sensitive to the spectral center of gravity of auditory stimuli (Roberts, et al., 2000). In the Fujioka et al. study, however, unfortunately the conditions are confounded, and therefore any differences in auditory evoked latencies could be attributed to significant differences in the spectral center of gravity. Therefore, to control for differences in the spectral center of gravity, while systematically modulating the induced fundamental component, I synthesized sinusoidal tone complexes with side bands that were kept constant across the different tokens (1200 Hz and 2400 Hz) and two additional sinusoids within these sidebands. This allowed us to systematically control the spectral center of gravity, while the internal sinusoids contributed the frequency of the inferred fundamental.

Methods and Materials

Participants

Nine (7 female; age range = 20-59; mean age = 26.3) healthy, right-handed adult volunteers with normal hearing participated in this study. All tested strongly right-handed on the Edinburgh Handedness Survey (Oldfield, 1971) and were compensated $10/hr for their participation. Each session lasted approximately 1½ to 2 hours. Participants provided written informed consent. The involvement of human participants in the reported experiment was approved by the University of Maryland, College Park Institutional Review Board (IRB).

Stimuli

Two different sets of auditory stimuli were synthesized using Praat (Boersma, 2001) at a sampling frequency of 44.1 KHz. Each stimulus was 70 ms in duration with 10 ms linear rise and decay ramps. The first set were pure sinusoids at 100 Hz, 200 Hz, 300 Hz, 400 Hz, 600 Hz, 1200 Hz and 2400 Hz. The second set of stimuli consisted of sinusoidal complexes. Each complex was composed of up to four component sinusoids. Two of the four sinusoids for all tone complexes were shoulder tones at 1200 Hz and 2400 Hz; the two other sinusoids were placed between the shoulder tones. The frequency of these two internal sinusoids varied to produce inferred fundamentals corresponding to the frequencies of the pure tone sinusoids. For example, the tone complex with an inferred fundamental component of 400 Hz was composed of equal amplitude sinusoids at 1200 Hz, 1600 Hz, 2000 Hz and 2400 Hz. One additional complex contained only the shoulder tones (i.e., 1200 Hz and 2400 Hz). The amplitudes of the sounds were chosen as a compromise between matching the physical sound level and the psychophysical intensity (i.e., from a hearing threshold curve). The complex stimuli had an average intensity of 84 dB SPL, and the pure sinusoids had an average intensity of 90 dB SPL, these values appeared to be relatively well-matched for listeners.

The particular nature of the structure of the tone complexes is important. First, by placing shoulder tones at 1200 Hz and 2400 Hz and successively moving the internal tones closer to the midpoint (i.e., 1800 Hz) in 100 Hz steps, we ensured that the spectral center of gravity (the first spectral moment, M_1) would remain constant across the tone complexes. This is evident in Table 3.1, where it is shown that the spectral center of gravity, M_1, is 1800 Hz across all tone complexes.

F_{Inf}	H_1	H_2	H_3	H_4	M_1	M_2	M_3	M_4
Pure sinusoids								
	100				100	8.82	0.82	37
	200				200	9.14	0.78	65
	300				300	9.24	0.76	94
	400				400	9.30	0.75	123
	600				600	9.35	0.74	179
	1200				1200	9.40	0.73	337
	2400				2400	9.43	0.72	606
Tone Complexes								
100	1200	1300	2300	2400	1800	552	-0.000020	-1.97
100	1200	1700	1900	2400	1800	430	0.000020	-1.10
200	1200	1400	2200	2400	1800	510	-0.000010	-1.85
300	1200	1500	2100	2400	1800	474	-0.000002	-1.64
400	1200	1600	2000	2400	1800	447	0.000009	-1.36
600	1200	1800	1800	2400	1800	490	-0.000002	-1.50
1200	1200			2400	1800	600	-0.000002	-2.00

Table 3.1: Spectral-values of the auditory stimuli.

F_{Inf} = Inferred Fundamental (in Hz); H_1 = First Harmonic (in Hz); H_2 = Second Harmonic (in Hz); H_3 = Third Harmonic (in Hz); H_4 = Fourth Harmonic (in Hz); M_1 = Spectral Centre of Gravity (in Hz); M_2 = Standard Deviation (in Hz); M_3 = Skewness; M_4 = Kurtosis

Again, this is important given that the latency of the M100 has been found to be sensitive to this property of the stimulus (Roberts, et al., 2000), a potential confound in some of the previous electrophysiological studies on the perception of the inferred fundamental (e.g., Fujioka, et al., 2003). Constructing the sinusoidal complexes in this manner also controls for skewness (the third moment, M_3) and kurtosis (the fourth moment, M_4). Thus, we can be confident in attributing the response profile of the M100 of these tone complexes solely to the inferred fundamental and not to some overall spectral shape property of the stimuli. Figure 3.1 presents a spectrogram showing all seven four tone complexes.

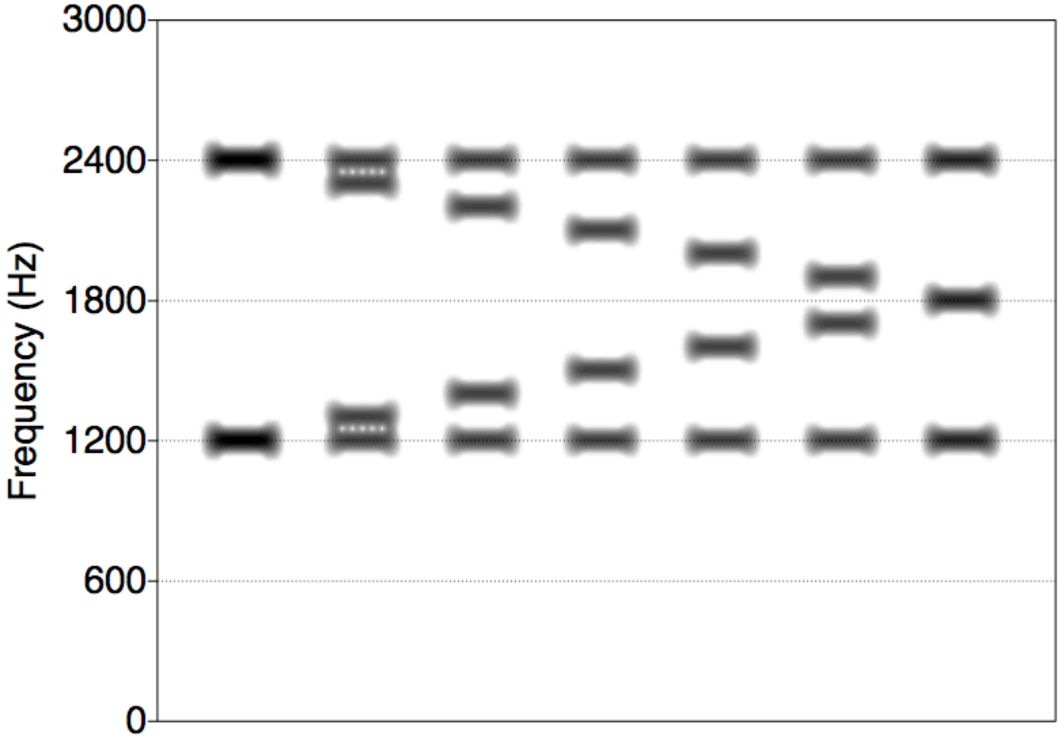


Figure 3.1: A composite spectrogram of the seven complex tones.

The duration of each complex tone was 70 ms, including 10 ms rise and decay time. Each complex tone included shoulder tones of 1200 Hz and 2400 Hz. Internal sidebands were synthesized in 100 Hz steps inward from the shoulder tones in six of the seven stimuli to induce the inferred fundamental components.

Procedure

Magnetoencephalographic recordings were made using a 157-channel whole-head axial gradiometer MEG system (Kanazawa Institute of Technology, Kanazawa, Japan). Participants lay supine in a magnetically shielded room. Auditory stimuli were delivered binaurally via Etymotic ER3A insert earphones. Earphones were calibrated to have a flat frequency response between 50 Hz and 3100 Hz within the shielded room. The inter-stimulus interval (ISI) varied pseudo-randomly between 700 ms and 1500 ms. All auditory stimuli were presented 150 times each. Stimulus-related epochs of 700 ms (200 ms pre-trigger) were averaged according to stimulus type to improve the signal-to-noise ratio. The neuromagnetic signal was sampled at 1 KHz with an online 200 Hz LPF and 60 Hz notch filter. Offline, the data were noise reduced using a multi-shift PCA noise reduction algorithm (de Cheveigné & Simon, 2007) and was band-pass filtered by a Hamming-window digital filter with frequency cut-offs at 0.03 Hz and 14 Hz. For each complex and each pure tone (corresponding to the missing fundamental), the same five source and five sink channels from the magnetic contour map that provided the strongest detected signal were selected from each hemisphere (20 total channels). M100 latency was defined as the root-mean-square (RMS) peak across these channels within a post-stimulus window of 90-180 ms and recorded, along with field strength (measured in fT), for each stimulus type. A 70 ms burst of broadband noise was presented as part of a distracter task. The noise burst was presented independently, occurring 150 times at pseudo-random intervals over five blocks of approximately 9 minutes.

Results

Figure 2 illustrates the RMS of a typical neuromagnetic response to both the pure sinusoid and its corresponding tone complex. Figure 3 shows mean M100 latency as a function of the fundamental frequency or missing fundamental. Statistical analyses were done using mixed-effects ANOVAs with Subject as a random effect, excluding the 12-17-19-24 complex tone to maintain a balanced design. Analysis of the latencies of the M100 responses showed main effects of frequency ($F(5,88) = 11.15$; $p < 0.0001$) and signal type (pure sinusoid vs. tone complex; $F(1,88) = 6.00$; $p = 0.016$), but crucially, there was no interaction between signal type and frequency ($F(5,88) = 1.02$; $p = 0.41$). In planned post-hoc comparisons, I found no significant differences at each frequency between the M100 latency to the pure sinusoid and the M100 latency to the tone complex, though the difference between the M100 response latency to the 100 Hz sinusoid and the 100 Hz inferred pitch tone complex (12-13-23-24) was marginally significant ($t(8) = 2.48$; $p = 0.015$, n.s. due to multiple-comparisons correction, all others $p > 0.12$). Analysis of the M100 amplitudes revealed a weakly significant main effect of frequency in which higher frequencies have larger amplitudes ($F(5,88) = 2.79$; $p = 0.022$), no main effect for signal type ($F(1,88) = 0.54$; $p = 0.46$), and a significant interaction between frequency and signal type ($F(5,88) = 5.82$; $p < 0.0001$). Post-hoc comparisons (Tukey-Kramer honestly significant differences) found only one significant contrast: the amplitude of the sinusoidal 100 Hz response is significantly weaker than the amplitude of the sinusoidal 1200 Hz response. The significant interaction effect is due to a cross-over between the sinusoidal responses (which have increasing amplitudes with increasing frequency) and a generally level amplitude response to all of the tone complexes.

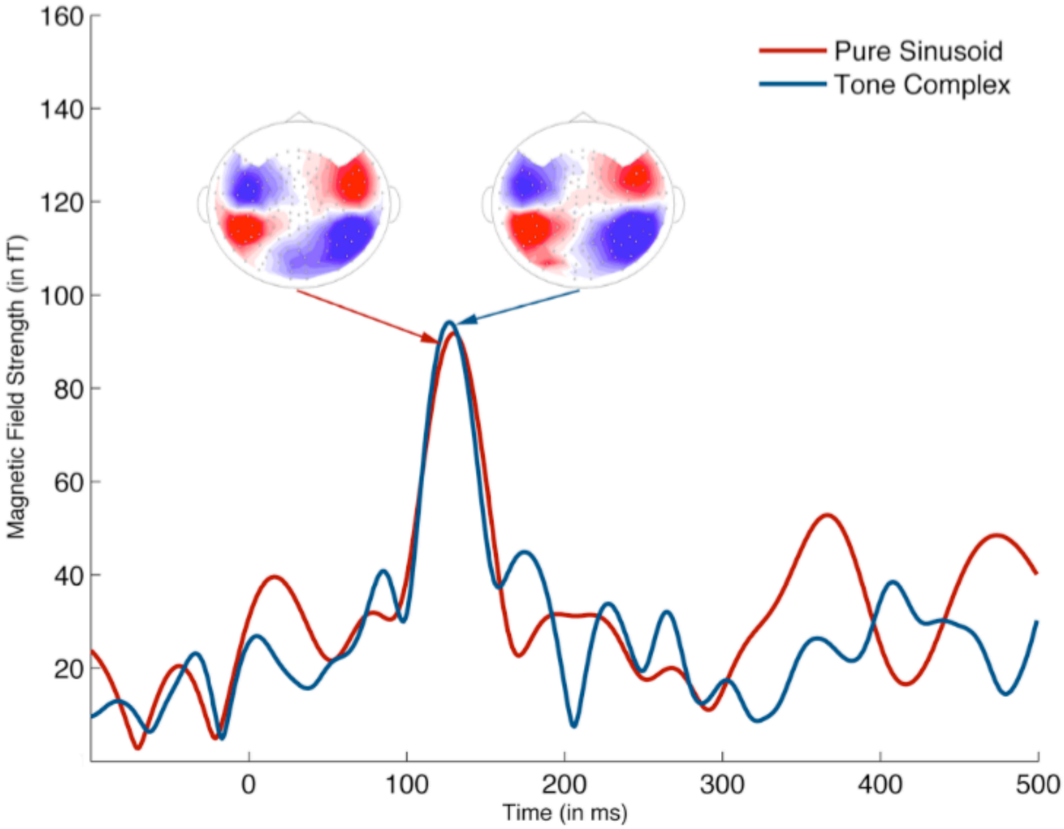


Figure 3.2: Overlay of RMS Waveforms

Comparison of the MEG waveforms to a pure sinusoid (in this case, 600 Hz) and tone complex with the corresponding inferred fundamental (in this case, 12-18-24) for a representative subject. Data is the RMS from 10 channels (five sink, five source) in the left hemisphere. The peak around 100 ms post-onset of the target (0 ms represents the onset of the target) is the M100. The peak latency of the M100 to the pure sinusoid and its corresponding tone complex were closely matched. The head-models represent the magnetic field contours for the M100. The red regions represent the source of the dipole and the blue regions represent the sink of the dipole.

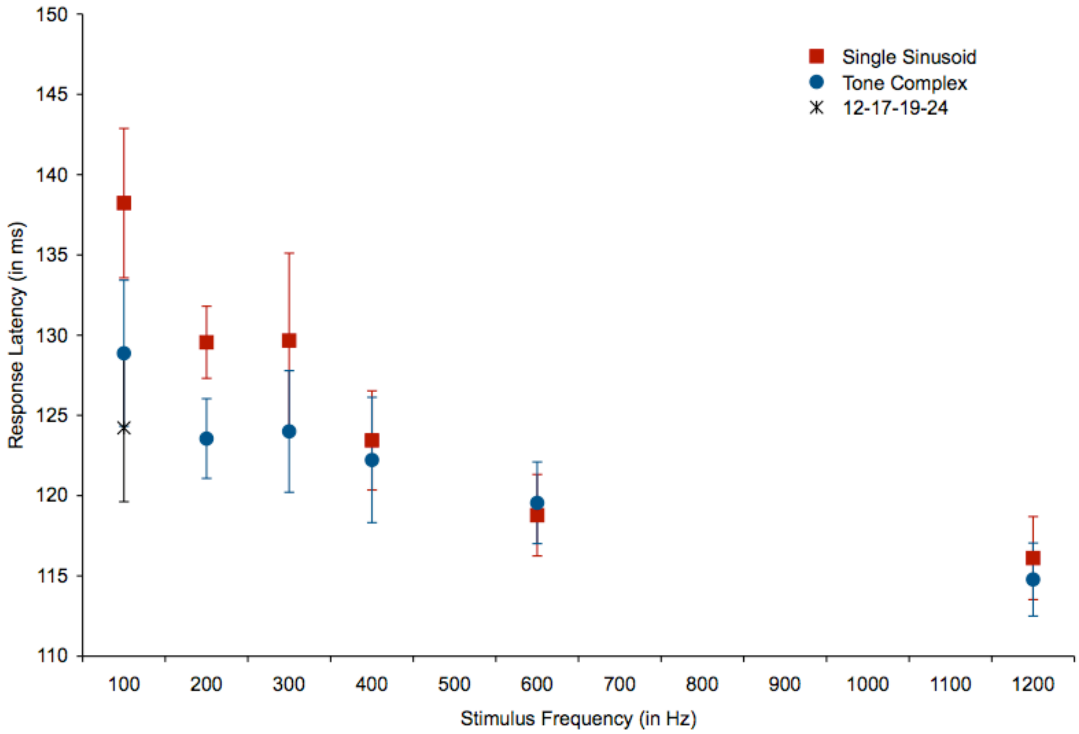


Figure 3.3: M100 RMS Latencies as a Function of Stimulus Frequency.

M100 RMS latencies to single sinusoid tones, tone complexes (plotted by their inferred fundamental component), and the 12-17-19-24 kHz tone complex, whose fundamental component is 100 Hz. Error bars refer to ±1 standard error of the group mean.

On a model that supposes that the M100 reflects just the physical properties of the stimulus, I would expect that the latencies to all tone complexes to be around 115 ms (the latency of the M100 to the 12-24 tone complex). In other words, I would anticipate that the 1200 Hz component present in each tone complex to drive a considerably faster M100 latency. This, however, is not the case. Instead, our findings suggest that the M100 is reflecting contributions of the inferred pitch of the stimulus and not solely the surface properties of the stimulus.[10]

[10] At the lower frequencies, there was a trend of the tone complexes eliciting slightly shorter M100 latencies than the pure sinusoids. This could be a result of the contribu-

Discussion

Using stimuli that incorporate a specific improvement over earlier materials, I replicated the M100 latency curve previously found (Roberts & Poeppel, 1996). Moreover, I found no latency difference between M100 responses to pure sinusoids versus tone complexes across frequencies. Our findings suggest that listeners are reconstructing the inferred pitch by roughly 100 ms after stimulus onset and are consistent with previous electrophysiological research suggesting that the inferential pitch is perceived in early auditory cortex (Matsuwaki, et al., 2004; Fujioka, et al., 2003; Pantev, Hoke, Lütkenhöner, et al., 1989; I. Winkler, et al., 1997). Moreover, the nature of the stimuli in the present study suggest that it is not necessary for a tone complex to be comprised of adjacent harmonics for pitch to be inferred (cf., Matsuwaki, et al., 2004).

These results provide information about the relative timing of when listeners reconstruct inferred pitch. In other words, whatever computations are germane to inferred pitch must be carried out in the initial stages of auditory processing. Understanding the time course of the perception of inferred pitch helps us to delimit the types of neurobiological computations involved. These findings do not allow us to decide between differing models of inferential pitch, but they do suggest that any model of pitch perception must place this reconstruction effect early in auditory processing. This conclusion is consistent with recent modeling research that proposes sub-

tion of the 1200 Hz component in the tone complexes. I do not believe it to be the case that the latency of the M100 is entirely dependent on the resolved fundamental component, but the physically present properties of the stimulus must also exert some effect on the electrophysiological component (though apparently not as strongly as the inferred pitch – otherwise, we should have found no differences between the various tone complexes).

cortical involvement in the reconstruction of virtual pitch via coordinated processing in populations of neurons (Meddis & O'Mard, 2006), which is what MEG measures. Research on the integration time of the M100 shows that the M100 integrates over the first 40 ms of signal (Gage & Roberts, 2000; Gage, et al., 1998; Gage, et al., 2006); therefore the computations we are seeing here must be executed over no more than that amount of input (see Chait, Poeppel, & Simon, 2006 for discussion of the spatial and temporal dynamics of pitch perception using MEG).

In addition to new information about inferred pitch, this study yields further insight into the nature of the M100 response itself. M100 latencies recorded in this study have been shown to co-vary with stimulus frequency when the stimuli were pure sinusoids, just as they were in Roberts and Poeppel (1996); but they have also been shown to vary with the inferred fundamentals of tone complexes. It is possible, then, to build on the findings in Roberts and Poeppel (1996) and conclude that the M100 reflects computations that are performed over the whole spectrum of the signal, and not simply an index of the transparent spectral properties of a stimulus.

Conclusion

MEG results suggest that listeners reconstruct the fundamental component of a complex tone early in auditory perception. In particular, by the time the neural generators of the M100 have been activated, I find evidence that listeners have reconstructed the fundamental component, indicating that some amount of abstract computations have been performed, in this case, the restoration of the fundamental component, early in auditory perception.

Chapter 4: Early Auditory Sensitivity to Formant Ratios[*]

Introduction

The perceptual and biological computations responsible for mapping time-varying acoustic waveforms onto linguistic representations remain far from understood (R. H. Fitch, Miller, & Tallal, 1997; Sussman, 2000; Hickok & Poeppel, 2007; Scott & Johnsrude, 2003). Speech includes not only the linguistic content of an utterance but also cues that allow listeners to infer sociolinguistic and personal information about the speaker (Ladefoged & Broadbent, 1957). These cues that allow listeners to recover speaker specific attributes, however, also serve to introduce significant acoustic variation, obscuring any straightforward one-to-one mapping of acoustic feature onto any linguistic representations that may exist.

Some of the most compelling demonstrations of this variability have been presented in acoustic analyses of vowel distributions across different talkers (Peterson & Barney, 1952; Potter & Steinberg, 1950). Given their tractable nature and well-understood spectral properties, vowels have played a central role in understanding the mechanisms that underlie speaker variation and normalization (Rosner & Pickering, 1994). The primary acoustic characteristic of spoken vowels is formants. Formants are the resonant frequencies of particular vocal tract configurations superimposed on

[*] Submited as:

Monahan, P.J., & W.J. Idsardi (submitted). Early Auditory Sensitivity to Formant Ratios: Toward a Perceptual Account of Vowel Normalization. *Language and Cognitive Processes*.

the harmonic resonances of the glottis pulse rate during production (Fant, 1960). Within speakers, the first (F1) and second (F2) formants are the primary determinants of vowel type—F1 varies as a function of vowel height and F2 varies as a function of vowel backness (F3 primarily cues rhoticity, Broad & Wakita, 1977).

While the relative pattern of formants remains constant across speakers for a given vowel type (Potter & Steinberg, 1950), the absolute formant frequencies for a given vowel vary as a function of vocal tract length (Huber, Stathopoulos, Curione, Ash, & Johnson, 1999). Using magnetic resonance imaging of the vocal tract, W. T. Fitch & Giedd (1999) demonstrate that vocal tract length positively correlates with age: children's' vocal tract lengths are significantly shorter than adults', though there is no difference between pre-pubescent boys and girls, and the average vocal tract length of post- pubescent males' is significantly longer than post-pubertal females'. Despite these significant differences in vocal tract length, listeners are quite good at recognizing phonemes across a number of different speakers (Strange, Jenkins, & Johnson, 1983; Smith, Patterson, Turner, Kawahara, & Irnio, 2005). Additionally, listeners are quite adept at estimating speaker size from modulations of vocal tract length with a minimal amount of speech information (Smith & Patterson, 2005; Ives, Smith, & Patterson, 2005). Taken together, these findings suggest that the auditory system employs special/specific mechanisms to solve the speaker normalization problem that allows listeners to robustly identify phonemes in the face of significant variation, as well as use information present in the signal to accurately identify attributes of the speaker. In other words, auditory cortex segregates the incoming speech signal into information that allows listeners to recover both the vocal tract size (formant scales) and vocal tract shape (formant ratios) contemporaneously with one another (Smith, et al., 2005; Irino & Patterson, 2002).

Recent functional neuroimaging and electrophysiological work has identified different cortical networks subserving the processing of speaker dependent ("who" is speaking) from speaker invariant ("what" is being said) features in vowel perception (Formisano, et al., 2008; Bonte, Valente, & Formisano, 2009). Specifically, Formisano, et al. (2008) showed that the cortical networks responsible for distinguishing vowel categories independent of speaker were more bilaterally distributed in superior temporal cortex and involved the anterior-lateral portion of Heschl's gyrus, planum temporale (mostly left lateralized) and extended areas of STS/STG bilaterally. The cortical networks underlying speaker identification independent of vowel category were far more right lateralized and included the lateral part of Heschl's gyrus (Heschl's sulcus) and three regions along the anterior-posterior axis of the right STS that were adjacent to areas in vowel discrimination. The conclusion is that neurobiology segregates the processing of speaker and vowel are consistent with recent perceptual learning (McQueen, et al., 2006; Norris, McQueen, & Cutler, 2003) and neurophysiological work (Obleser & Eisner, 2009) arguing that listeners must construct abstract pre-lexical representations of phonological categories that are independent of particular speakers and that episodic traces (Johnson, 2005; Pisoni, 1997; Johnson, 1997; Goldinger, 1996a; Bybee, 2001; Pierrehumbert, 2002) cannot be the only representational schema employed in speech perception.

As discussed above, listeners require little input to reliably judge both personal aspects of a speaker (Perry, Ohde, & Ashmead, 2001; Smith & Patterson, 2005;

Smith, et al., 2005; Ives, et al., 2005), and their sociolinguistic background (Purnell, Idsardi, & Baugh, 1999). Furthermore, pre-linguistic infants are able ignore speaker-dependent acoustic variation and successfully categorize vowels across different talkers (Kuhl, 1979, 1983). Taking all these findings into consideration, it appears that whatever normalization procedures are available to listeners are available without significant linguistic experience or exposure to novel speakers, and thus it seems that accumulating large amount of speaker-dependent information is unnecessary to adequately normalize across speakers. In order for a vowel normalization algorithm to be computationally plausible, it must satisfy at least these two conditions: 1) It must be shown that the algorithm adequately normalizes the vowel space, and 2) it must be shown that the human perceptual system is sensitive to the computations required by the algorithm in online speech perception.

The primary goal of this chapter is to revisit an idea that has received sporadic attention throughout the past in attempting to solve the speaker normalization problem: formant ratios (Miller, 1989; Syrdal & Gopal, 1986; Johnson, 2005; Lloyd, 1890; Peterson, 1951, 1961; Potter & Steinberg, 1950). In particular, we pursue a specific instantiation of formant ratios that has received little attention in the literature, namely that information in higher formants, specifically the third formant (F3), acts as the normalizing factor (Peterson, 1951; Deng & O'Shaughnessy, 2003). If this particular hypothesis is on the right track, then we are also proposing that the appropriate dimensions for the vowel space are the ratios F1/F3 and F2/F3 (or logarithmic-like transforms of these quantities, such as Bark difference scores) as opposed to the traditional F2 by F1 vowel space. We calculated the extent to which this particular hypothesis (F1/F3 by F2/F3) removed inter-speaker variation based on age and gender of the talkers from the Hillenbrand, Getty, Clark, & Wheeler (1995) corpus of American English vowels. Subsequently, we present data from two magnetoencephalographic (MEG) experiments that suggest that auditory cortex is sensitive to modulations of the F1/F3 ratio. Additionally, our findings indicate that the perceptual system displays heightened sensitivity to formant ratios in more densely populated regions of the vowel space (in English, this would be for front and back vowels and not mid vowels).

Formant Ratios

Potter and Steinberg (1950) and Peterson and Barney (1952) were the first to demonstrate how poorly vowels clustered together along the traditional dimensions of the vowel space (F1 and F2). That is, when the first and second formant frequencies of ten American English vowels as spoken by men, women and children were plotted along the F2 by F1 axes, there was no clear separation between different vowel types. This famous plot is reproduced in Figure 4.1.

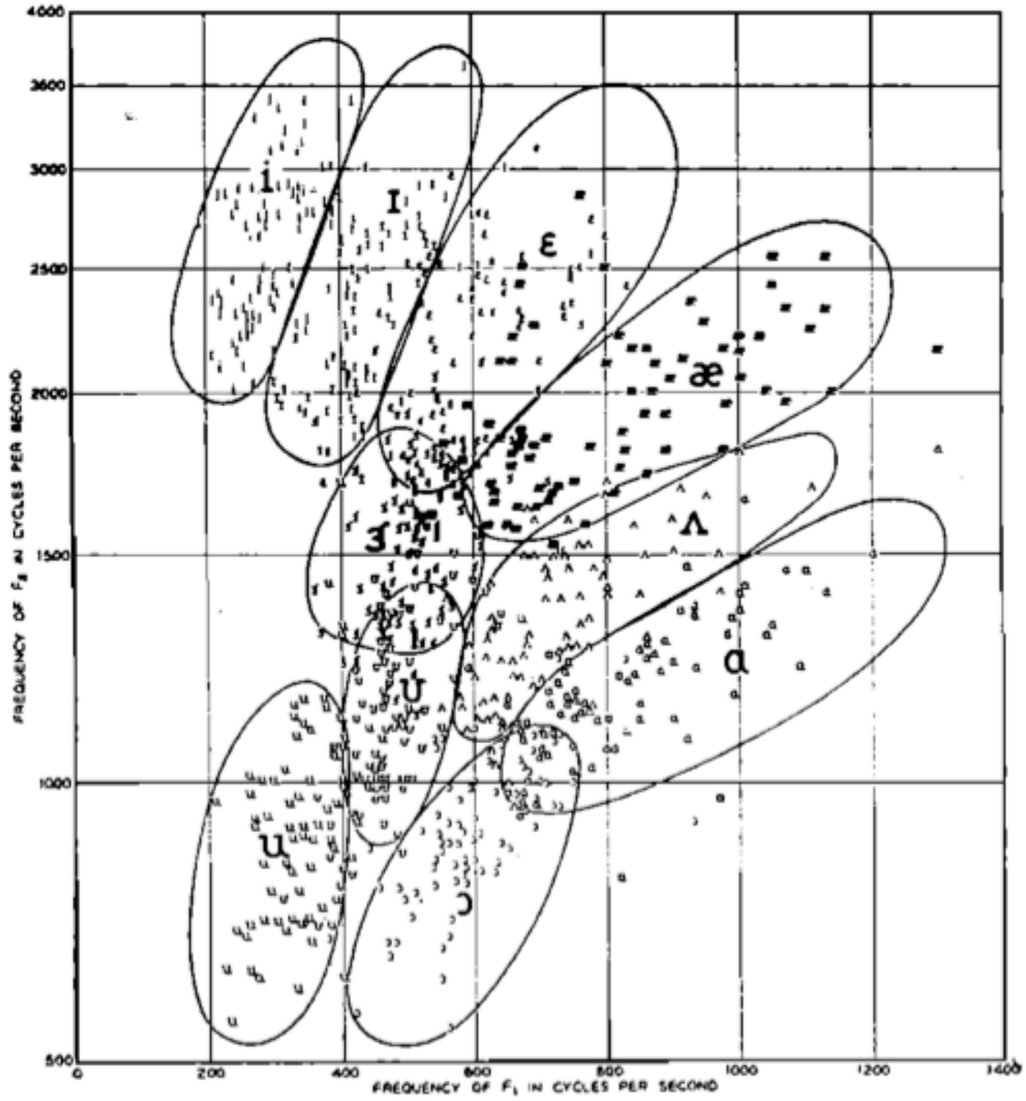


Figure 4.1: Scatter Plot of Vowel Tokens in American English.

One of the first explicit demonstrations of the amount of overlap between vowel tokens of different categories when produced by different speakers (76 total; 33 men, 28 women, 15 children) across a variety of vowel categories. [Reprinted from Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184; © 1952 *Journal of the Acoustical Society of America*]

This led to the conclusion that a simple mapping from F2 by F1 vowel space to vowel categories was insufficient. In other words, listeners could not rely solely on distributions in an F2 by F1 vowel space to adequately categorize vowel tokens. Therefore, the primary problem, then, in vowel normalization has been to understand the set of mechanisms, computations and acoustic cues that allow listeners map this highly variable acoustic input onto invariant representations.

The notion of the formant ratio has been present in the literature for over a century (Lloyd, 1890) but has never received consistent attention (Miller, 1989). The idea for formant ratios is quite simple: listeners are sensitive to the *relative* differences between and not the absolute values of the formants themselves. Vowels of the same quality should exhibit homologous formant patterns across individuals, less the specific frequencies (see Johnson 2005 for additional discussion and criticism). While Potter and Steinberg (1950), Peterson (1951) and Peterson and Barney (1952) were the first to use the idea of formant ratios in trying to solve the normalization problem, Peterson (1961) was the first to devote space to the problem. In his particular algorithm, he calculated the log difference (subtraction in log space is equivalent to division in linear space; hence, log differences are ratios; i.e., log(a)-log(b)=log(a/b)) between the second, third and fourth formants and the first ($\log F_n - \log F_1$ (n=2,3,4)). He found that for the triplet [ɪ], [ʌ] and [ɯ], the vowels can either be distinguished on the basis of the first two formants alone or by the formant ratios. Unfortunately, his particular ratio hypothesis faired worse on the triplet [ɛ], [œ] and [ɝ]. Peterson (1961) concludes that neither the absolute value of the formants nor a formant ratio hypothesis are adequate for “explaining vowel perception fully” (p. 24), and consequently suggests that fundamental frequency, amplitude and phonetic environment must also be considered. We certainly do not disagree with the latter claim. But, the algorithm proposed in Peterson (1961) uses F1 as the normalizing factor, and given that F1 is one of the primary determinants of vowel type and varies as a function of vowel height and tongue root placement, it seems that a more reliable normalizing factor could be used instead.

In a more recent approach, Syrdal and Gopal (1986) present a model of vowel perception that attempts to understand the mechanisms involved in removing speaker-dependent acoustic variance. At an intermediate stage of perceptual processing, Bark difference scores are computed between the higher formant and its lower neighbor and between the first formant and the fundamental frequency (Bark(F1)-Bark(f0), Bark(F2)-Bark(F1), Bark(F3)-Bark(F2)). For Syrdal and Gopal (1986), the F1-F0 dimension is responsible for classifying vowel height. Front vowels have a F1-F0 difference less than three Bark, while mid and low vowels have a Bark difference greater than three Bark. The frontness dimension is characterized by the F3-F2 dimension. Back vowels have a Bark difference greater than three, while front and central vowels have a Bark difference less than three. The Bark differences of these transformed vowels are then sent to a subsequent phonetic stage of processing, where categorization is performed. Vowels whose difference scores are greater than three Bark are classified as being distinct. The vowel tokens were a subset of the Peterson and Barney (1952) data. The model proposed in Syrdal and Gopal (1986) did a fair job at correctly classifying the different vowel tokens that were either within or greater than 3 Bark different (total correct: high: /i/ and /ɝ/, 99.3%; low: /ʊ/, 84%). In this case there is no single normalizing factor; rather adjacent spectral prominences are compared. It is not immediately clear in this case whether one of the extracted quantities yields a reasonable assessment of the speakers' vocal tract lengths.

Finally, Miller (1989) presented a normalization procedure not dissimilar from Syrdal and Gopal (1986). He took the ratios between adjacent formants in the logarithmic scale: [log(F3/F2)], [log(F2/F1)], [log(F1/SR)]. SR refers to the sensory reference, which is essentially a given speaker's fundamental frequency over some determined time window. Plotting the results in three-dimension auditory-perceptual space, Miller (1989) demonstrated that his algorithm performed with 93% accuracy. Accuracy is defined as how well the classification algorithm correctly categorized the vowel tokens into their pre-determined, intended vowel category.

A handful of criticisms have been leveled against formant-ratio theories of vowel perception. Strange (1989) defined two broadly held positions in the vowel perception literature. The first is the "Elaborated Target Approaches" view. In this approach, vowels are processed via spectral measurements taken from a particular point in the steady state. Strange (1989) argues that formant ratio theories fall into this first category. Crucially, for critics, no information regarding the trajectory of the formant patterns is considered (Hillenbrand & Nearey, 1999; Hillenbrand & Gayvert, 1993). The second approach, the "Dynamic Specification Approach" considers the dynamic trajectories of the formants. Zahorian & Jagharghi (1993) performed automatic vowel classification experiments and demonstrated that in the absence of fundamental frequency information, classification algorithms that encode spectral shape properties fair better in categorization across a variety of contexts than those that simply encode target formant value. Adank, Smits & van Hout (2004) compared twelve vowel normalization procedures on how well they categorize Dutch vowels, while eliminating the physiologically-based variation and retaining the sociolinguistic information embedded within the speech signal. They found that normalization procedures that used vowel-extrinsic information (characteristics across multiple vowels) faired better than vowel-intrinsic procedures. Moreover, normalization procedures that operated on multiple formants (the ratio of F3/F2, for example) were not as accurate as procedures that used information from only one formant. While this work has highlighted the importance of considering the trajectories of formant structures, it appears that the importance of the steady-state values of formants has not been dismissed (Strange, 1989; Nearey, 1989). The particular hypothesis we explore in this chapter is vowel-intrinsic in nature, and while we discuss it here in terms of an "Elaborated Target Approach", we see no reason why it cannot incorporate more dynamic properties of vowel formants. The advantage of this algorithm (F3 as the normalizing factor) is that it requires no prior experience with the specific talker, which is consistent with the perceptual evidence described above that suggests that listeners do not need large amounts of exposure to adequately normalize vowels, and inconsistent with classification algorithms, which seem to require large amounts of corpus data in order to correctly classify vowels. The model here considered could easily be extended to include information from trajectories in F1/F3 by F2/F3 space, incorporating the thereby transformed results from the "Dynamic Specification Approach."

Third Formant Normalization

It is clear that F2 by F1 is not an adequate model of the vowel space (Peterson & Barney, 1952). The proposal put forward in this article is that vowel categorization is accomplished via normalizing the first and second formants against the third. Consequently, we propose that, representationally, the vowel space is best viewed not as F2 plotted against F1 or the difference between F2 and F1 plotted against F1, but rather as the ratio of F1 to F3 plotted against the ratio of F2 to F3. As a result, F3 acts as the normalizing factor. Given that the third formant (F3) is useful in the identification and discrimination of a variety of speech contrasts (rhoticization on vowels (Broad & Wakita, 1977), /l/-/r/ distinction (Miyawaki, et al., 1975), stop consonant place of articulation identification (Fox, Jacewicz, & Feth, 2008)), we can be confident that listeners are able to use and exploit information contained within this frequency range. Figure 4.2 illustrates the distribution of mean F3 across ten vowel categories for men, women and children of American English.

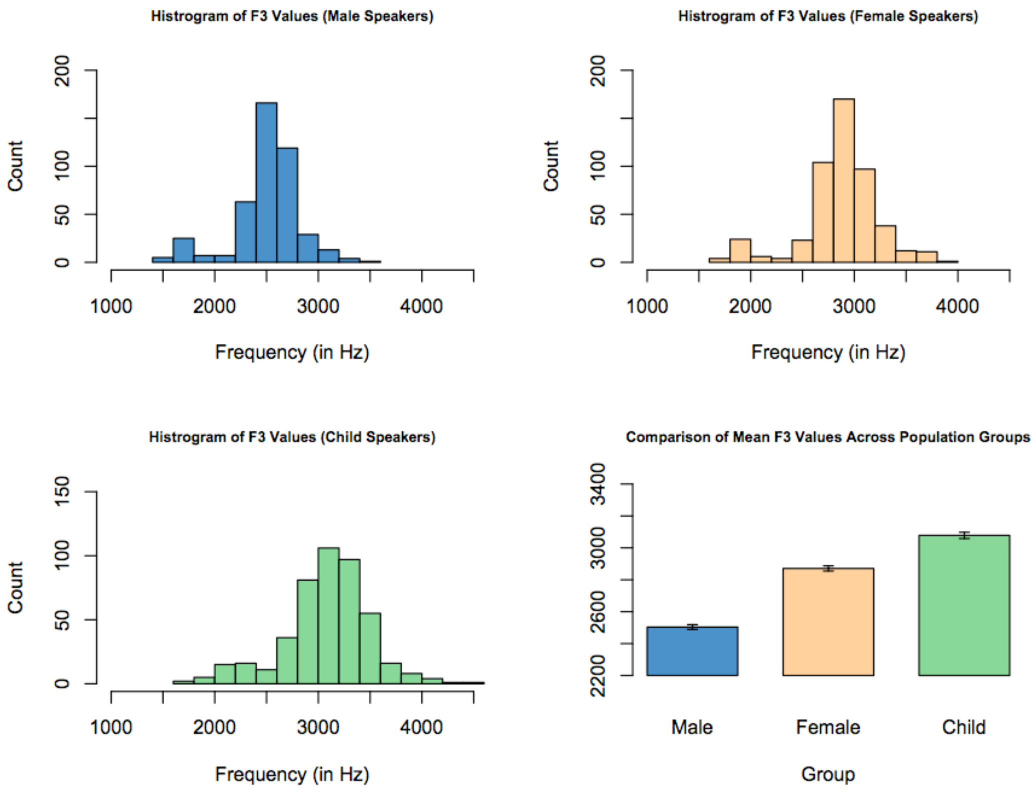


Figure 4.2: Distribution and Mean values of F3 by Speaker Group

Distribution and mean values of the third formant (F3) by speakers of American English broken down by group (men, women, children). Values reported are taken from (Hillenbrand, et al., 1995). Notice the scaling of the raw Hz value of F3 dependent upon speaker group: the mean F3 for men is lowest (2504 Hz), then F3 for women (2871 Hz) and the highest mean F3 for children (3078 Hz). This scaling is presumably a consequence of vocal tract length (W. T. Fitch & Giedd, 1999).

Additionally, it should be pointed out that listeners are not likely mapping the raw frequencies in Hz onto their perceptual space but are instead performing auditory transformations on the input signal, such as BARK (Zwicker, 1961), Mel (S. S. Stevens & Volkmann, 1940), Koenig (Koenig, 1949), etc., which are much more appropriate scales based on psychophysical performance (Mel) or the auditory system (BARK). Many of these scales transform much of the frequency space into log space, and therefore, computing differences between formants in these scales is largely equivalent to computing the ratios of the formants in a linear space.

Numerous previous formant ratio proposals have included f0 into their algorithm. Fundamental frequency has been shown to correlate with speaker dependent factors such as gender, age and size. We have three reasons for attempting to pursue an algorithm that does not include f0, however. First, whispered speech is intelligible and lacks a fundamental frequency (Kiefte, 2005). Presumably, we want our normalization algorithm to work for both phonated and whispered speech. Second, numerous languages, such as Japanese, Korean, Comanche and Southern Paiute employ vowel devoicing (lack of phonation/f0), either as an allophonic process or as part of their phonological inventory. Presumably, devoiced vowels should also be normalized. Inclusion of f0 in the normalization algorithm would require additional mechanisms to normalize these vowels. Third, tone languages, such as Vietnamese, Mandarin Chinese and Yoruba, use f0 to encode the linguistic tone present on a vowel. Large contour modulations render the inclusion of f0 in our normalization algorithm questionable. We, instead, suggest that F3 is a more promising solution to the normalization problem.

Delattre, Liberman, Cooper, & Gerstman (1952) synthesized vowels containing only F1 and F2 from a pattern playback machine. They presented the synthesized vowel tokens to listeners and found that they were quite identifiable, suggesting that perhaps only the first two formants are necessary for vowel identification. These findings minimally demonstrate is that the perceptual system can adequately deal with two formant vowel tokens if that is all that is provided. In other words, it can manage with degraded stimuli and categorize them appropriately when only given degraded stimuli. Additionally, we know from Peterson & Barney (1952) and Potter & Steinberg (1950) that the spread of F2/F1 frequencies for a given vowel type spoken by many different individuals are quite diffuse, and moreover, there is considerable overlap between adjacent categories. The implicit assumption in this comparison is that categorization on F2/F1 planar space is homologous to speaker characterization of a given token. What is suggested then, is that F1 and F2 may be sufficient for identification of synthesized vowels, but for categorization across numerous natural tokens by numerous speakers, more than F1 and F2 is required. This sentiment was echoed in Peterson & Barney (1952) and Potter & Steinberg (1950). Broad & Wakita (1977) suggested this be the case in order to appropriately categorize rhoticized vowels, such as [ɚ] in ‘writer’ [ɹɐj.ɾɚ]. The third formant, F3, seems to be the critical energy band in rhoticized vowels.

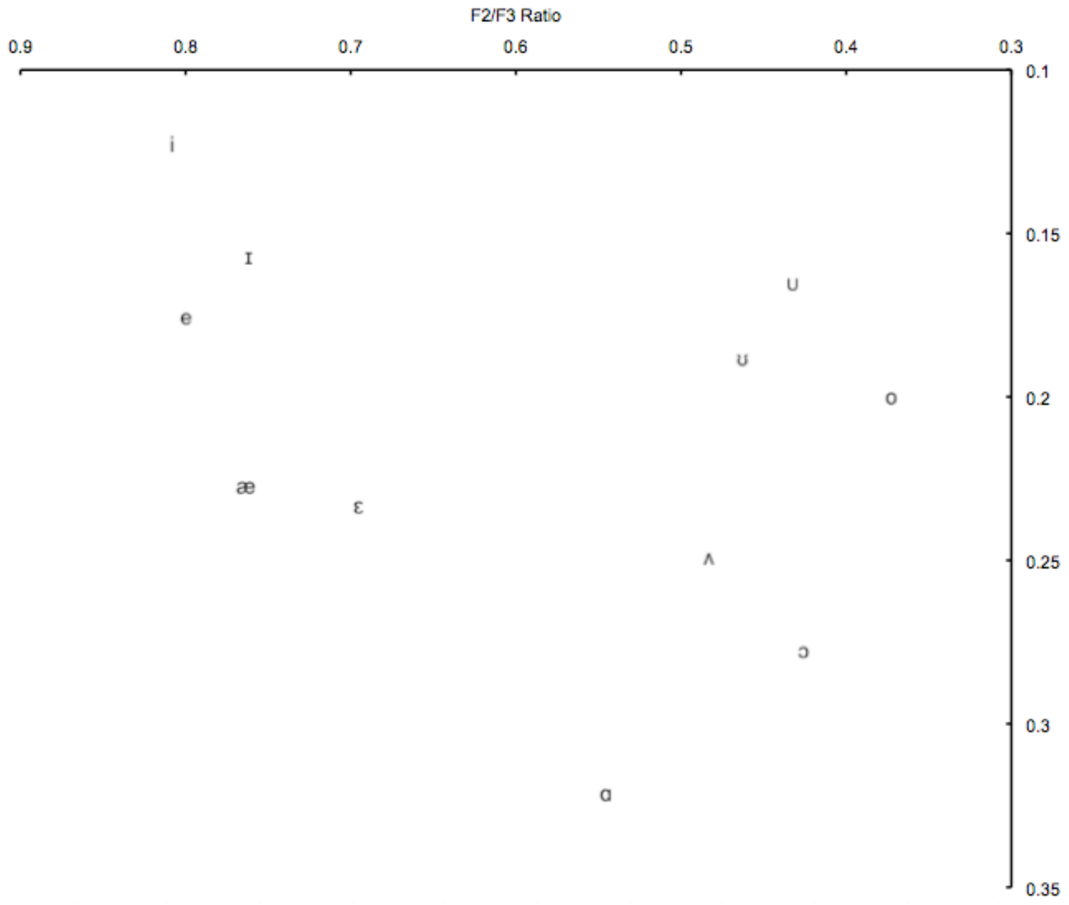


Figure 4.3: Vowel Space Normalized Against F3.

Traditional vowel space plotted in the proposed normalized vowel space (F3 as the normalizing factor). Formant values from which ratios were computed are from Hillenbrand, et al. (1995) and averaged across age and gender per vowel category. Variation in font size is not meaningful.

The frequency of F3 for a given vowel has been shown to vary correlationally with the fundamental frequency of the speaker and holds fairly constant across vowels (with the exception of rhoticized vowels) for a given speaker (Potter & Steinberg, 1950; Deng & O'Shaughnessy, 2003), and F3 has been shown to effect the perception of vowels (Slawson, 1968; Fujisaki & Kawashima, 1968; Nearey, 1989). Given that F3 remains fairly constant across vowels for a given speaker, it might serve usefully in normalizing across multiple speakers. The third formant (F3) has also been shown to provide a good estimate of vocal tract length in automatic speech recognition (Claes, Dologlou, ten Bosch, & van Compernolle, 1998) and is useful in normalizing whispered vowels (Halberstam & Raphael, 2004; although their data on the role of F3 in normalizing phonated vowels was inconclusive). It has also been shown that the higher formants are as important, if not more so, than pitch in normalizing noise-excited vowels (Fujisaki & Kawashima, 1968) and that F3 . Potter & Steinberg (1950) converted formant frequencies into Mel space and plotted F2/F1 and F3/F2. They concluded from this exercise that ratios may be necessary but are not sufficient to remove inter-talker variability. This is because [ɔ] and [ɑ] could not accurately be categorized from one another nor could [u] and [ʊ]. This was one of the criticisms that Miller (1989) cites against formant ratio theories. That, in particular, vowel types of different categories have similar ratios. It has long been known, however, that the formant frequencies of F1 and F2 are intimately tied to vocal tract configuration. The first formant varies as a function of tongue frontness and F2 as a consequence of tongue height. Therefore, a ratio theory of inter-talker normalization where one of the two dimensions has both sides of a ratio varying as a function of vowel category seems less than ideal.

Given that the third formant appears to be relatively stable across vowel tokens within a given speaker, but varies as a function of vocal tract length inter-talker, and is present in different types of speech (e.g., whispered), an alternative solution would be to take the ratio of the first and second formants against the third. In the final two pages, Peterson (1951) did exactly that. He converted vowel frequencies into Mel space and plotted F1/F3 against F2/F3. Taking vowel productions from one man, one woman and one child, he showed that, impressionistically, this ratio removes much of the variation seen when F2 is plotted against F1. Unfortunately, little discussion or further results are provided, and it seems that this particular algorithm has not been pursued subsequently in the formant ratio literature. It should be pointed out, however, that a similar sentiment was echoed in Deng and O'Shaughnessy (2003, p. 252), where they write:

> “Since F3 and higher formants tend to be relatively constant for a given speaker, F3 and perhaps F4 provide a simple reference, which has been used in automatic recognizers, although there is little clear evidence from human perception experiments.”

This is the particular hypothesis of formant ratios and speaker normalization we investigate here. It should also be pointed out that one of half of the algorithm we propose (F2/F3) has been present in a previous algorithms (Syrdal & Gopal, 1986; Miller, 1989). Therefore, in our experiment, we concentrate on finding neurophysiological evidence for the more novel ratio, the F1/F3 ratio.

Computational Evidence

While the objective of this chapter is to demonstrate human perceptual sensitivity to formant ratios, it is useful to assess how well our proposed algorithm eliminates variance due to speaker differences. The corpus data used to test our model is from Hillenbrand, et al. (1995). In a replication of Peterson and Barney (1952), Hillenbrand, et al. (1995) collected the productions of twelve American English vowels from 45 men, 48 women and 46 children in an /hVd/ frame. In an acoustic analysis of the data, they identified a point centrally located in the steady-state portion of the vowel and measured the fundamental frequency (f0), as well as the center frequencies for the first through fourth formants (F1-F4) for each token.

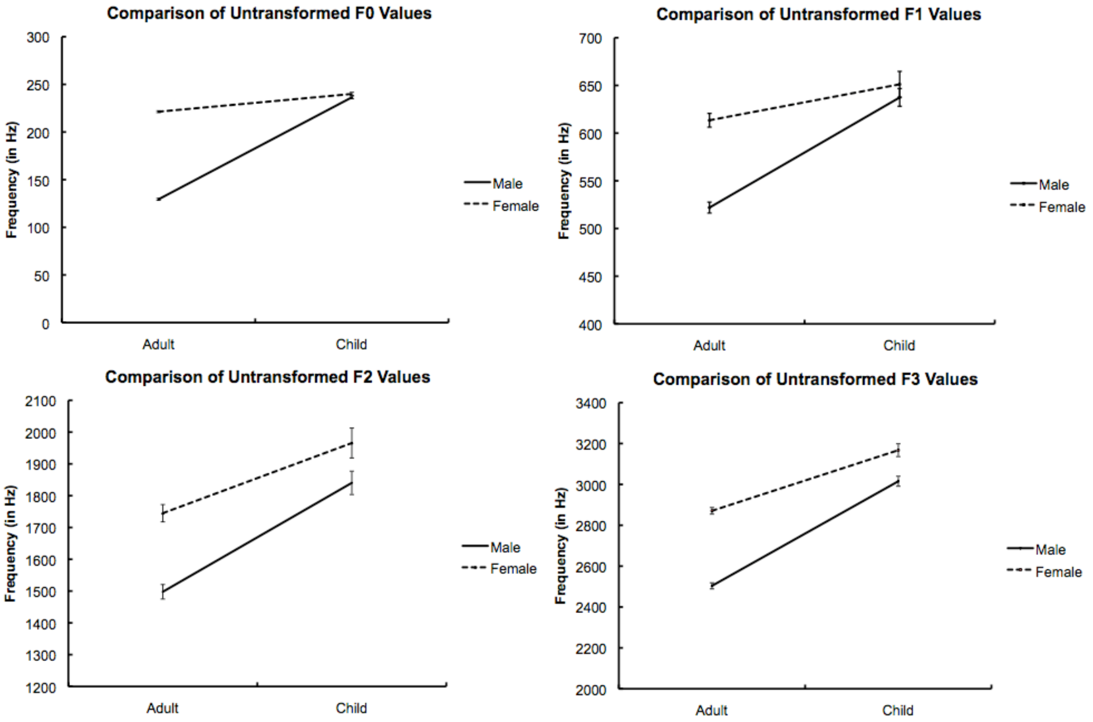


Figure 4.4: Comparison of Untransformed Mean Formant Values by Group.

A comparison of the untransformed formant values across all vowel categories for fundamental frequency (f0), first formant (F1), second formant (F2), and third formant (F3) by age and gender of speaker. Mean formant values by group calculated from Hillenbrand, et al. (1995).

In order to assess the amount of inter-speaker variation in the data as a function of speaker age and gender, we performed a Mixed Effects ANOVA with Subject as a Random Variable on the Hillenbrand, et al. (1995) data comparing the effects of Age ('adult', 'child') and Gender ('male', 'female') on the raw frequency values for f0-F3 and subsequently on the transformed F1/F3 and F2/F3 ratios. For f0, we found a significant main effects for both Age (F(1,1612) = 2278.1, $p < 0.0001$) and Gender (F(1,1579) = 1540.3, $p < 0.0001$), and a significant Age × Gender interaction (F(1,1577) = 1560.8, $p < 0.0001$). For F1, we found a significant main effect of both Age (F(1,629.2) = 117.6, $p < 0.0001$) and Gender (F(1,1400) = 55.6, $p < 0.0001$) and a significant Age × Gender interaction (F(1,1449) = 18.97, $p < 0.0001$). For F2, we again found a significant main effect of Age (F(1,652.4) = 127.98, $p < 0.0001$) and Gender (F(1,1391) = 34.71, $p < 0.0001$) and a signification Age × Gender interaction (F(1,1430) = 6.54, $p < 0.02$). Finally, for F3, we once more found a significant main effect of both Age (F(1,938.5) = 450.79, $p < 0.0001$) and Gender (F(1,1582) = 174.80, $p < 0.0001$) and a significant Age × Gender interaction (F(1,1599) = 20.96, $p < 0.0001$). These results confirm that the values of formants across men, women and children are highly variable, and effects of age and gender contribute to this variability. The conclusion to draw from these findings is that a simple mapping between frequency information and speaker-independent representations is inadequate.

To provide at least an initial demonstration how well our proposed algorithm eliminates speaker variation, we ran the same model above on the transformed (F1/F3; F2/F3) corpus data. The significant main effects of age and gender, as well as the significant interactions of age and gender were largely eliminated. For F1/F3, there were no main effects of Age (F(1,599.9) = 0.09, $p = 0.76$) or Gender (F(1,1366) = 0.51, $p = 0.47$) and only a marginal Age × Gender interaction (F(1,1410) = 2.85, p

= 0.09). For F2/F3, we still found a main effect of Age (F(1,860.6) = 6.17, *p* < 0.02), but no main effect of Gender (F(1,1472) = 0.58, *p* = 0.44), and no Age × Gender interaction (F(1,1479) = 0.82, *p* = 0.36).

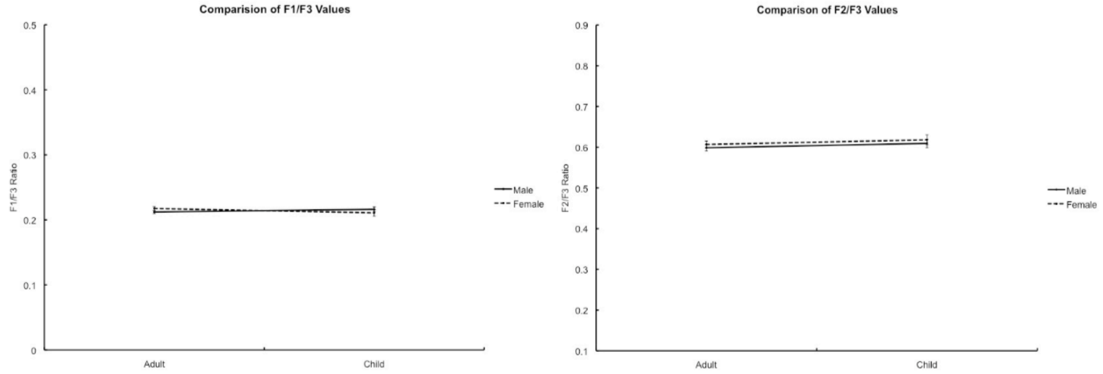


Figure 4.5: Comparison of Mean Transformed Formant Values by Group.

A comparison of the transformed values for F1/F3 and F2/F3 by age and gender of speaker. Normalized formant values calculated from Hillenbrand, et al. (1995).

The results of Mixed Effects ANOVA on the transformed Hillenbrand, et al. (1995) data suggests that our proposed algorithm, whereby F1 and F2 are ratioed against F3 successfully eliminates most of the variance due to effects of age and gender found in productions of vowel tokens across different speakers. The question we now focus on, and the primary aim of this chapter, is to demonstrate that auditory cortex is sensitive to one of the two dimensions of our proposed formant ratio algorithm, namely F1/F3. To do this, we present data from two MEG experiments on vowel perception, and our findings confirm that auditory cortex appears to be sensitive to modulations of the F1/F3 ratio (note that the F1/F3 ratio is the more novel of the two dimensions, as a host of previous algorithms have adopted F2/F3 as one of the dimensions).

The Contribution of Magnetoencephalography

Magnetoencephalography (MEG) is an electrophysiological recording technique that measures fluctuations in magnetic field strength caused by the electrical currents in neuronal signaling (Hari, Levänen, & Raij, 2000; Lounasmaa, Hämäläinen, Hari, & Salmelin, 1996) and is particularly adept at recording potentials from auditory cortex (Roberts, et al., 2000). Combining its excellent temporal resolution (1 ms; and fair spatial resolution ~2-5 cm) and aptitude for recording from auditory cortex, it provides a powerful tool in understanding how humans process speech in real time, whose temporal properties are both fast and fleeting. The MEG temporal waveform has been shown to ignore irrelevant stimulus variation (f0, Voice Onset Time (VOT) within category). In the mismatch negativity (MMN, MMF, MMNm) paradigm (Näätänen, 2001), listeners habituate to a series of standards that vary along some irrelevant parameter for categorization, which is interrupted by a deviant stimulus that differs from the standard along some dimension (e.g., category boundary). If the perceptual system is sensitive to this difference, a mismatch field is elicited roughly between 150 ms and 300 ms post onset of the deviant (Aulanko, et al., 1993; Phillips, et al., 2000; Kazanina, et al., 2006).

In the two experiments reported below, however, we exploit the response latency of an earlier evoked magnetic potential, the M100 (or N1m), which is the MEG equivalent of the N1 ERP component (Virtanen, et al., 1998; Eulitz, et al., 1995). The dependent measures of the evoked M100 (latency, amplitude) more directly reflect properties of the acoustic stimulus (frequency, loudness, fine-structure of the wave-form, etc.), as opposed to later evoked components (e.g., MMNm), and integrates only over the first 40 ms of the auditory stimulus (Gage, et al., 2006). Given its robustness and replicability, the M100 has been used extensively to study early auditory cortical processing, and we have a fair understanding of the types of stimulus dependent factors to which the M100 is sensitive (Roberts, et al., 2000). Sinusoids closest to 1 KHz elicit the shortest evoked M100 response latency, while moving outward from 1 KHz in either direction (both lower and higher in frequency) elicit longer evoked latencies (Roberts & Poeppel, 1996). Relevant to the current work, the M100 response properties to vowels have been fairly well characterized. In particular, the M100 seems to be sensitive to F1 (Govindarajan, et al., 1998; Roberts, et al., 2000; Poeppel, et al., 1997; Hannu Tiitinen, et al., 2005; Diesch, et al., 1996; Roberts, et al., 2004) independent of differences in fundamental frequency (Govindarajan, et al., 1998; Poeppel, et al., 1997). Diesch, et al. (1996) compared the evoked latencies of the M100 to four different synthesized German vowels (/a/, /i/, /u/, /æ/) and found that /a/ and /æ/, having higher F1 values, elicited reliably shorter latencies than /u/. Poeppel, et al. (1997) synthesized three English vowels (/i/, /u/, /a/) and also report a reliable difference in the evoked M100 latency between /a/ and /u/, with /a/ eliciting a shorter M100 latency, and do not report a difference between /i/ and /u/. This finding was replicated in Govindarajan, et al. (1998), who also found that both one and three formant synthesized tokens of /a/ elicit reliably faster M100 evoked latencies than one and three formant synthesized tokens of /u/, respectively, in English listeners.

Tiitinen, et al. (2005) replicate these findings in Finnish speakers, showing again that /a/ elicits faster M100 latencies than /u/ using semi-synthetic speech. The interpretation for the directionality of these effects, namely that /a/ elicits reliably shorter M100 evoked latencies than /u/, is that the spectral energy in F1 is driving the M100 response (Govindarajan, et al., 1998; Poeppel, et al., 1997; Roberts, et al., 2000), and /a/ elicits shorter latencies because the F1 in /a/ (~ 700 Hz) is significantly closer to 1 KHz than the F1 in /u/ (~ 300 Hz), consistent with the sinusoidal data (Roberts & Poeppel, 1996). This effect does not seem to be speech specific, however (Diesch, et al., 1996; Govindarajan, et al., 1998). These findings have been confirmed and extended in more recent work. Roberts, et al. (2004) show that unlike the sinusoidal data, where the M100 response latency follows a smooth 1/f(requency) function (at least through 1 KHz, above 1 KHz the latency again increases), the latency of the M100 to vowels (F1, in particular) seem to respect vowel category boundaries. They synthesized tokens of /a/ and /u/ and modulated F1 in 50 Hz increments between 250 Hz and 750 Hz while keeping the values for F2 (1000 Hz) and F3 (2500 Hz) consistent across the vowel tokens, albeit with bandwidths that were broader than usual. Instead of following the smooth 1/f function that the sinusoids elicit, the M100 latencies clustered into three distinct bins, the lowest F1 values (250 – 350 Hz) elicited the longest latencies, the middle F1 values (400 – 600 Hz) elicited reliably shorter latencies and the high F1 values (650 – 750 Hz) elicited even shorter M100 latencies. The bin with the lowest F1 values also represent the natural range of F1 in /u/ and the bin with the highest F1 values represent the natural range of F1 in /a/ tokens. Therefore, Roberts, et al. (2004) concluded that the latency of the M100 is sensitive to information about the F1 frequency distributions of different vowel categories. Collectively, these data suggest that the M100 is sensitive to F1 in vowel perception. Vowel categories with a higher F1 (closer to 1 KHz) consistently elicit shorter evoked latencies of the M100.

Experiment 1

The goal of this chapter is to determine if the perceptual system is sensitive to formant ratios, and in particular if the perceptual system is sensitive to the F1/F3 ratio (one of the two dimensions in our proposed normalization algorithm). Given our hypothesis regarding the algorithm that is (at least partly) responsible for vowel normalization, combined with the previous MEG findings on vowel perception (Govindarajan, et al., 1998; Roberts, et al., 2000; Poeppel, et al., 1997; Hannu Tiitinen, et al., 2005; Diesch, et al., 1996; Roberts, et al., 2004), we propose that the M100 is sensitive to the ratio of the first formant (F1) against the third (F3) instead of just F1 alone. In order for us to test this representational and normalization hypothesis with the M100, the M100 must be able to index more complex auditory operations performed on the input and not solely reflect surface properties of the stimulus. The results from Roberts, et al. (2004) and work on inferential pitch perception that has shown that the M100 is modulated by a missing fundamental component (Monahan, et al., 2008; Fujioka, et al., 2003) demonstrate that the M100 can index more complex and abstract auditory operations which integrate information from across the spectrum.

In the first experiment, participants were with tokens of [ɛ] and [ə], holding F1 (and F2) constant while manipulating the value of F3 for each type. We modulated F3 both higher and lower by 4% in Mel space from the mean/standard F3 value (8% overall difference between the two tokens for a given vowel type). We predict that vowels with a lower F3 (larger F1/F3 ratio) should elicit faster M100 latencies than vowels with a higher F3 value (smaller F1/F3 ratio). This directional prediction is derived from recalculating the formant values of Poeppel, et al. (1997). Converting their vowel tokens (for the male fundamental frequency) into F1/F3 Mel space, we find that the token of /a/ has a larger F1/F3 ratio than the token of /u/ and these two tokens are 20% from each other in this transformed space. Given that /a/ elicited a shorter M100 latency than /u/ (Govindarajan, et al., 1998; Roberts, et al., 2000; Poeppel, et al., 1997; Hannu Tiitinen, et al., 2005; Diesch, et al., 1996; Roberts, et al., 2004), we therefore expect tokens with a larger F1/F3 ratio to elicit shorter M100 latencies than tokens with a smaller F1/F3 ratio.

Participants

Thirteen monolingual English participants (5 female; mean age: 20 yrs old) participated in the experiment. For one participant, the stimuli did not elicit a reliable M100, and this participant was excluded from the analysis. Consequently, data from 12 participants was analyzed. Participants reported no hearing deficits. All participants provided written informed consent approved by the University of Maryland Institutional Review Board (IRB) and scored strongly right-handed on the Edinburgh Handedness

Survey (Oldfield, 1971). Each participant was compensated $10/hour. The typical session lasted approximately 1 ½ to 2 hours.

Materials

Vowel tokens were synthesized using HLSyn (K. N. Stevens & Bickley, 1991). Two tokens for each vowel type (mid-vowels /ɛ/ and /ə/) were synthesized, for a total of four tokens. A fundamental frequency of 150 Hz was used for all tokens. Using a f0 halfway between a male and female speaker allowed for greater flexibility in our possible F3 values. Moreover, a fundamental frequency of 150 Hz is not outside the possible range of either male or female speakers.

As previously mentioned, the values of F1 and F2 remained consistent across the tokens within each type. The Hertz values were converted into Mel space and we modulated the F3 value 4% higher and 4% lower in frequency space. Each token was 250ms in duration with a 10 ms $\cos^2$ on- and off-ramp. The values for F1, F2 and F3, and their respective bandwidths are presented in Table 4.1 and a comparison of the LPC-based spectral envelopes of the vowel tokens are presented in Figure 4.6.

		F1		F2		F3	
Vowel Type	F3 Height	Center Frequency	Bandwidth	Center Frequency	Bandwidth	Center Frequency	Bandwidth
/ə/	Low	500	80	1500	90	2040	150
/ə/	High	500	80	1500	90	3179	150
/ɛ/	Low	580	80	1712	90	2156	150
/ɛ/	High	580	80	1712	90	3247	150

Table 4.1: Experiment 1: Spectral characteristics of the four vowel tokens

The center frequency and bandwidth for each of the first three formants are provided in Hertz. The stimuli were synthesized using KLSyn (K. N. Stevens & Bickley, 1991), a user interface for the HLSyn speech synthesizer. The formant ratio calculations were performed in Mel frequency space and then converted back into Hertz for the speech synthesis.

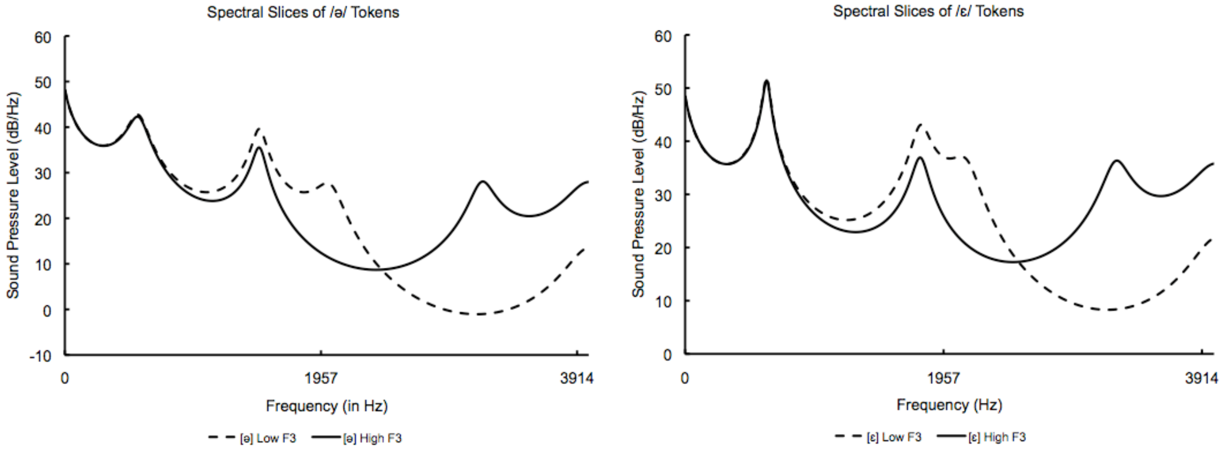


Figure 4.6: Spectral slices of the two /ə/ tokens and two /ɛ/ tokens in Experiment 1.

LPC-based spectral envelopes of the vowel sounds used in Experiment 1. The solid line indicates the token with a higher F3 (smaller F1/F3 ratio) and the dashed line indicates tokens with a lower F3 (larger F1/F3) ratio. Spectral envelopes have been smoothed with a six pole LPC filter.

The F1, F2 and F3 values for [ə] are standard values (K. N. Stevens, 1998). The F1 and F2 values (and the F3 value for which we computed from) for [ɛ] are taken from a recent corpus of American English vowel formant frequencies (Hillenbrand, et al., 1995) extracted from the steady state portion of the vowel in [hVd] syllables. For present purposes, we used the average formant values for male speakers.

Procedure

Participants lay supine in a dimly lit magnetically shielded room, as stimulus evoked magnetic fields were passively recorded by a whole-head 157-channel axial-gradiometer magnetoencephalography system (Kanazawa Institute of Technology, Kanazawa, Japan). The stimuli were delivered into the magnetically shielded room via Etymotic ER3A insert earphones, which were calibrated and equalized to have a flat frequency response between 100 and 5000 Hz. Prior to the experiment, a hearing test was administered to the participants within the MEG system to ensure normal hearing and that the auditory stimuli were appropriately delivered by the earphones. Subsequently, a pretest localizer was performed. Participants listened to roughly 100 tokens each of four pure sinusoids: 125Hz, 250Hz, 1000Hz and 4000Hz. The neuromagnetic-evoked responses to the sinusoids were epoched and averaged online. The pretest was done to ensure good positioning of the participant’s head within the system, as well as guaranteeing that the participant would show a reliable M100 response. The experiment began subsequent to the hearing test and pretest localizer.

For the experiment itself, participants listened to both vowel tokens and pure sinusoids. The four vowel tokens (/ə/: high F3; low F3; /ɛ/: high F3; low F3) were each presented 300 times (1200 vowel tokens in total), ensuring a good signal-to-noise ratio in the MEG signal. Sinusoids of 250Hz and 1000Hz were played 50 times each throughout the experiment. Participants were asked to listen passively to the vowel tokens and discriminate between the 250 Hz and 1000 Hz sinusoids. The inter-trial interval pseudo-randomly varied between 700ms and 1300ms.

Recording and Analysis

Neuromagnetic signals were acquired in DC (no high pass filter) at a sampling frequency of 1 KHz. An online Low Pass Filter of 200 Hz and a 60 Hz notch filter were applied during recording. Noise reduction was performed on the MEG data using a multi-shift PCA noise reduction algorithm (de Cheveigné & Simon, 2007). During the averaging process, any trials with artifacts exceeding 2.5 pT in amplitude during their epoch were removed from the analysis (6.2% of the total data). Off-line filtering (digital Band Pass Filter with a Hamming window, range: 0.03 - 30 Hz) and baseline correction (100 ms prior to onset of the vowel) were performed on the averaged data. Ten channels from each hemisphere that best correlated with the sink (ingoing magnetic field; 5 channels) and source (outgoing magnetic field; 5 channels) of the signal were selected for statistical analysis on a participant-by-participant basis. The same channels were used across the four conditions for the within subjects analysis. The peak latency and amplitude of the root mean square (RMS) of the evoked M100 in the MEG temporal waveform for each hemisphere were carried forward for statistical analysis.

Results

Given that we had a specific prediction regarding the direction of the effect in the latency of the RMS of the M100 (Larger F1/F3 ratios should elicit shorter latencies), we report statistics from one-tailed tests (statistics on the amplitudes are two-tailed).

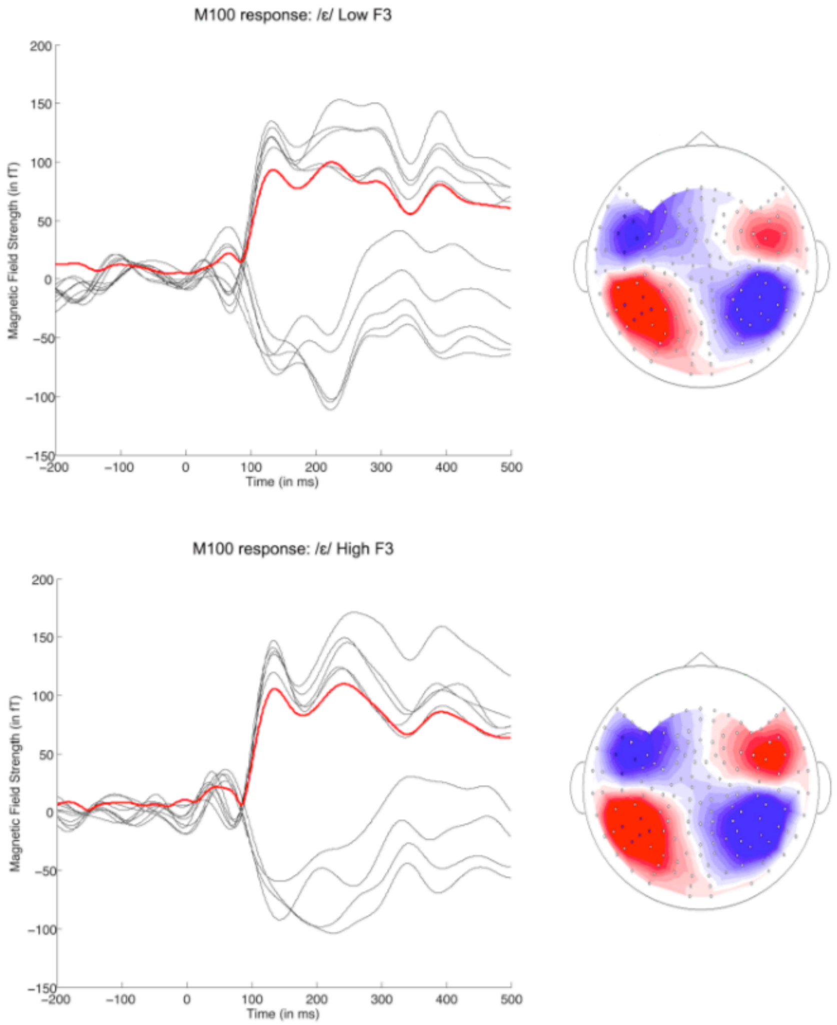


Figure 4.7: Evoked M100 Temporal Waveform and Magnetic Field Contour for Representative Subject.

Temporal waveform from ten left hemisphere channels (RMS solid red line superimposed) and the magnetic field distribution at peak latency of the M100. Waveform and field distribution to /ɛ/ token with Low F3 (larger F1/F3 ratio) presented along the top row and waveform and field distribution to /ɛ/ token with High F3 (smaller F1/F3 ratio) presented along the bottom row. For the magnetic field distributions, red indicates outgoing magnetic field source and blue indicated ingoing magnetic field sink.

Pooling across both hemispheres (20 channels), we found a reliable difference in the peak latency of the RMS of the M100 for the vowel /ɛ/ (paired t-test; t(11) = 1.996, $p < 0.05$). As predicted, the /ɛ/ token with the lower F3 (larger F1/F3 ratio; M

= 134 ms) elicited a significantly shorter M100 latency than the /ɛ/ token with the higher F3 (smaller F1/F3 ratio; M = 138 ms). Analyzing the hemispheres independently, we found a reliable difference in the right hemisphere (RH; t(11) = 2.412, *p* < 0.02) and a marginal effect in the left hemisphere (LH; t(11) = 1.753, *p* = 0.054).

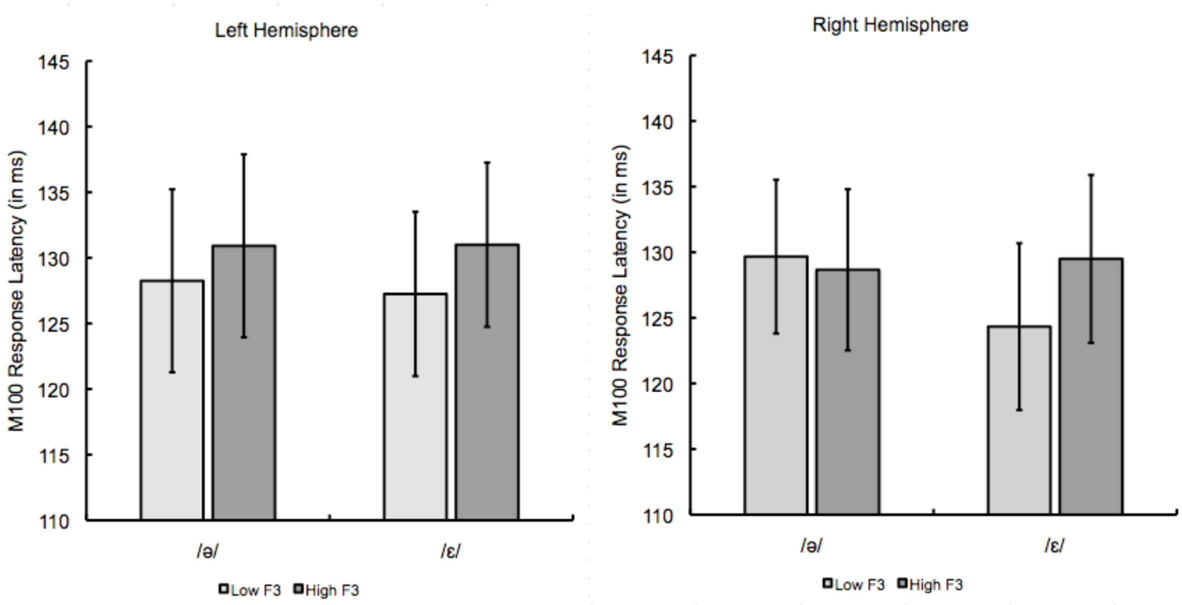


Figure 4.8: M100 Response Latencies by Vowel Type.

Mean M100 response latencies across participants to the vowel tokens in Experiment 1. Light gray bars refer to tokens with a Low F3 (large F1/F3 ratio) and dark gray bars refer to tokens with a high F3 (small F1/F3 ratio). Error bars represent one standard error of mean.

We find the same pattern of effects when we perform the non-parametric equivalent statistic (Wilcoxon Signed-Rank Test; Both Hemispheres: V = 64, *p* < 0.05; LH: V = 62, *p* < 0.05; RH: V = 68, *p* < 0.02). We found no reliable differences in the amplitude of the RMS of the M100 in the evoked temporal MEG waveform between the two tokens of /ɛ/ (Both Hemispheres: t(11) = 1.172, *p* = 0.13; LH: t(11) = 0.158, *p* = 0.44; RH: t(11) = 0.540, *p* = 0.30).

Comparing the two tokens of /ə/, we did not obtain reliable differences in the latency of the RMS of the M100 response between the /ə/ token with a large F1/F3 ratio (M = 136 ms) compared with the /ə/ with a small F1/F3 ratio (M = 138 ms) when we pooled sensors across both hemispheres (t(11) = 0.996, *p* = 0.17), nor did we find a difference between the two tokens of /ə/ when analyzing sensors from just the left hemisphere (t(11) = 0.983, *p* = 0.17) or the right hemisphere (t(11) = -0.387, *p* = 0.65) in the peak latency of the RMS of the evoked M100. Again, we did not find any differences in the peak amplitude of the RMS of the M100 to the two tokens of /ə/ when we pooled sensors across both hemispheres (t(11) = -1.303, *p* = 0.22), or in the left hemisphere (t(11) = -0.828, *p* = 0.43), though we did find a marginal effect in the right hemisphere (t(11) = -2.176, *p* = 0.053). To summarize, we found a reliable difference in the latency of the evoked M100 component in the predicted direction, vowel tokens with a larger F1/F3 ratio elicit a shorter M100 latency than tokens with a smaller F1/F3 ratio, which is consistent with our reinterpretation of the previous M100 results on F1 (Govindarajan, et al., 1998; Roberts, et al., 2000; Poeppel, et al., 1997; Hannu Tiitinen, et al., 2005; Diesch, et al., 1996; Roberts, et al., 2004). We did, however, only find this effect for the front vowel /ɛ/ and not the central-vowel /ə/. (It should be noted that central vowels have yet to be tested in the literature to the best of our knowledge.) In the discussion, we speculate on some potential explanations for this pattern of results. In general, however, we take these results to suggest that the perceptual system is sensitive to formant ratios, and the F1/F3 ratio in particular.

M100 ECD Source Location

The primary aim of this study was to determine if the auditory perceptual system is sensitive to formant ratios in general, and moreover, if manipulations of the F1/F3 ratio would modulate the response latency of the RMS of the evoked auditory M100 component. In addition to the latency and amplitude analysis of the M100, we also conducted a source analysis on the data to determine if there were any localization differences between the vowels. Obleser, Lahiri, & Eulitz (2004), using MEG, calculated the Equivalent Current Dipole (ECD) of the source for seven distinct German vowels. They found that front vowels tend to map onto a more anterior portion of auditory cortex while back vowels map onto a more posterior region of auditory cortex. Thus, the front/back distinction of vowel categories is retained on the anterior/posterior dimension of early auditory cortex. Provided these results, in our experiment, we might expect to find reliable differences between the tokens of /ɛ/ and /ə/, with the front vowel /ɛ/ localizing to more anterior regions than the mid vowel /ə/. Based on a sphere model, we calculated the ECD solution of the four distinct vowel tokens on an intra-subject and intra-hemispheric basis (minimum GoF = 90%, mean GoF = 95.8%; one participant was excluded as we were unable to calculate a fit with a GoF > 90%, statistics on source n = 11). We performed a Repeated-Measures ANOVA with the factors Vowel (/ə/ and /ɛ/) and F3 ('High' and 'Low'). In the left hemisphere along the lateral-medial plane (*x*-axis; sagittal), we find no main effect of Vowel (F(1,10) = 0.179, p = 0.68) or F3 (F(1,10) = 0.346, p = 0.57) and no interaction of Vowel × F3 (F(1,10) = 0.433, p = 0.53). In the superior-inferior plane (*z*-axis; axial), we again find no main effect of Vowel (F(1,10)=0.023, p = 0.88) or F3

(F(1,10) = 1.510, p = 0.25) and only a marginal interaction of Vowel × F3 (F(1,10)=3.186, p = 0.10).

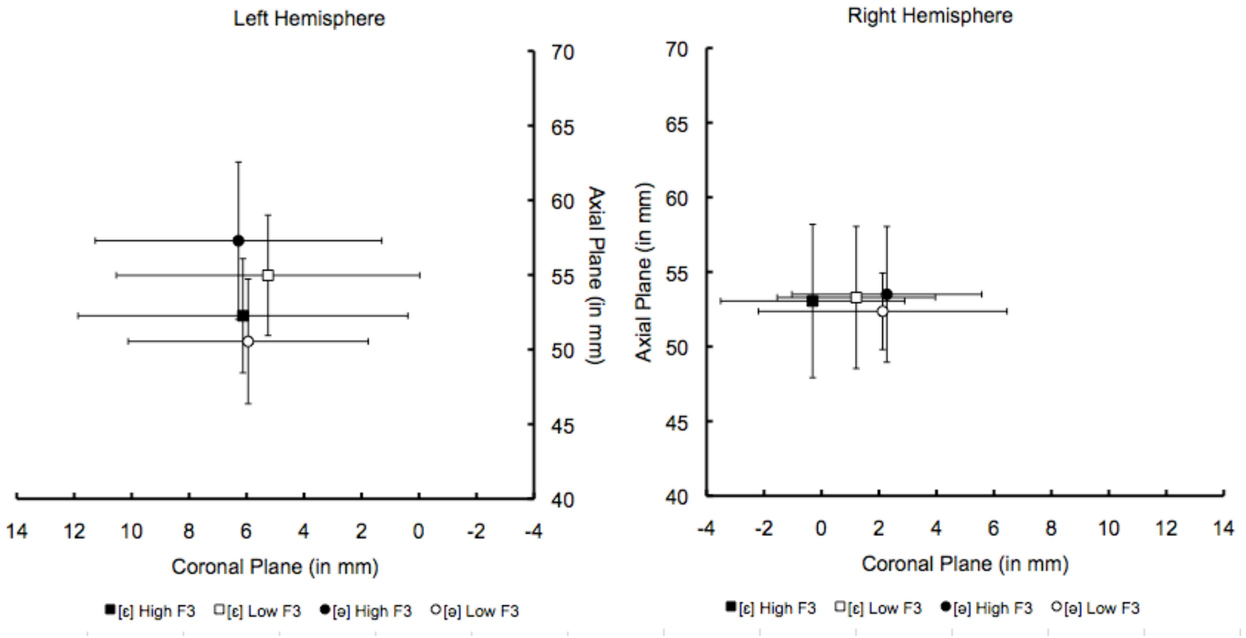


Figure 4.9: Mean ECD Locations along the Coronal (y) and Axial (z) Planes.

Plot of the mean ECD source locations for each vowel token in Experiment 1 plotted along the Coronal (y; anterior-posterior) and Axial (z; superior-inferior) planes. Error bars represent one standard error of the mean.

Finally, along the inferior-posterior dimension (*y*-axis; coronal), where we might expect to find a difference given Obleser, et al. (2004), we also find no main effect of Vowel (F(1,10) = 0.072, p = 0.79) or F3 (F(1,10) = 0.16, p = 0.70) and no interaction of Vowel × F3 (F(1,10) = 0.02, p = 0.89). The final statistic we calculated was to determine whether the front vowels (the two tokens of /ɛ/) were located more anterior than the mid-vowels (the two tokens of /ə/). We performed a one-tailed sign test on the values of y (the anterior-posterior plane) across vowel type for the left hemisphere, and we found no difference between the tokens with a high F3 (S = 6; *p* = 0.5) or between the tokens with a low F3 (S = 4; *p* = 0.89), implying that we were unable to replicate the anterior-posterior localization difference reported in Obleser, et al. (2004).

In the right hemisphere, for the lateral-medial plane (*x*-axis; sagittal), we again find no main effect of Vowel (F(1,10) = 1.565, p = 0.24) or F3 (F(1,10) = 2.192, p = 0.17) and no interaction of Vowel × F3 (F(1,10) = 0.518, p = 0.49). In the superior-inferior (*z*-axis; axial) plane we also find no main effect of Vowel (F(1,10) = 0.047, p = 0.83) or F3 (F(1,10) = 0.176, p = 0.68) and no interaction of Vowel × F3 (F(1,10) = 0.66, p = 0.44). Finally, along the anterior-posterior plane, we again find no main effect of Vowel (F(1,10) = 2.072, p = 0.18) or F3 (F(1,10) = 0.293, p = 0.6) and no interaction of Vowel × F3 (F(1,10) = 0.384, p = 0.57). Performing a one-tailed sign test on the values for the anterior-posterior (*y*-axis; coronal) plane we find no difference between vowel types for the tokens with a high F3 (S = 4; p = 0.89) or with a low F3 (S = 4; p = 0.89).

Discussion

These findings suggest that auditory cortex (minimally the neurobiological generators of the M100) is sensitive to modulations of the F1/F3 ratio.[11] The latency difference was robust across participants for the /ɛ/ vowel. Recall that in the beginning of the chapter, we stated that in order for a vowel normalization algorithm to be computa-

[11] To be more confident in this finding, additional research is needed that shows that co-variation in F1 and F3 that does not modulate the F1/F3 ratio should not elicit latency differences in the M100.

tionally plausible, it must satisfy at least these two conditions: 1) It must be shown that the algorithm adequately normalizes the vowel space, and 2) it must be shown that the human perceptual system is sensitive to the computations required by the algorithm in online speech perception. The findings from the first experiment, at least, suggest a positive answer to the second condition for the particular algorithm we are testing. Our results suggest that the human auditory system is sensitive to formant ratios. The latency difference was in the predicted direction for nearly all subjects (Sign Test: $p < 0.05$) for the vowel type /ɛ/. Any more concrete conclusions, however, should be taken cautiously, given that we did find such an effect for the mid-central vowel /ə/.

The immediate question is why we found an effect of F3 manipulation for /ɛ/ but not for /ə/. The lack of a result for /ə/ is not likely due to a lack of power in the experiment, given that 300 tokens of each vowel is more than sufficient to obtain a good signal-to-noise ratio. Moreover, the fact that we found an effect with /ɛ/ suggests that this asymmetry is due to some intrinsic properties of the vowels or their location in vowel space. It is this second possibility that we explore in the second experiment. In particular, the asymmetry found in Experiment 1 could be a consequence of the location in vowel space that /ɛ/ and /ə/ occupy. The front mid-vowel /ɛ/ occupies a more crowded portion of the vowel space relative to that occupied by /ə/ (i.e., there are many more phonetic categories in close proximity to the distribution of /ɛ/ as opposed /ə/ in the vowel space), where categorization might be more critical than the middle of the vowel space.

In Experiment 2, we test the hypothesis that the asymmetry found in Experiment 1 is due to the location in vowel space of each vowel. Consequently, we test two hypotheses. First, we aim to replicate the null effect with /ə/ that we found in Experiment 1. To accomplish this, we test the same /ə/ tokens with a different set of participants. Second, to test whether it is the demands for categorization that drive the perceptual system's sensitivity to formant ratios in more crowded portions of the vowel space, we test two tokens of /o/ with the same manipulations we performed on /ɛ/ in the first experiment. The back vowel /o/, like /ɛ/, also occupies a more crowded portion of the vowel space than /ə/.

Experiment 2

The second experiment was nearly identical to Experiment 1; however, instead of testing tokens of /ɛ/, we tested synthesized tokens of /o/, a vowel produced in the back of the mouth. The back mid rounded vowel /o/, like /ɛ/, resides in a more crowded portion of the vowel space, at least when compared with the central vowel /ə/. Practically speaking, our hypothesis predicts we should find M100 latency differences for vowels located in more crowded portions of the vowel space. Therefore, we should find effects for /o/ and replicate our null effects for /ə/. We speculate that the reason for this particular pattern is likely due to greater competition within the category space, which drives the perceptual system's heightened sensitivity to the formant ratios in these more densely populated regions.

Participants

Fifteen monolingual English participants (9 female; mean age: 20 yrs old) participated in the experiment. Six participants were excluded from analysis on various grounds: two participants were not included in the analysis, as there were no discernable M100 in the data; two participants were excluded from the analysis because the source distribution of the component did not match that of an M100; one participant was excluded because the peak latency of their M100 was over 200 ms; and finally, one participant was excluded due to hardware failure. Consequently, for the analysis, the data from nine participants was analyzed. All participants had normal hearing. All participants provided written informed consent approved by the University of Maryland Institutional Review Board (IRB) and scored strongly right-handed on the Edinburgh Handedness Survey (Oldfield, 1971). Each participant was compensated $10/hour. The typical session lasted approximately 1 ½ to 2 hours.

Materials

For the /ə/ stimuli, we used the same tokens used in Experiment 1. For the /o/ stimuli, we synthesized two new tokens using HLSyn (K. N. Stevens & Bickley, 1991) with F1 and F2 values from (Hillenbrand, et al., 1995). Again, we converted the Hz frequency values into Mels. Using the F3 value (transformed into Mel space) from Hillenbrand, et al. (1995) as the standard, we computed the new F3 values for our experimental tokens by moving 4% in either direction of the F1/F3 ratio space. Therefore, the overall distance in F1/F3 ratio space between the tokens was 8%. As before, we predict an M100 latency facilitation for the token with the smaller F3 (the larger

F1/F3 ratio). The F1, F2 and F3 values for the four tokens used in Experiment 2 are presented in Table 4.2 and a comparison of the LPC-based spectral envelopes of the vowel tokens are presented in Figure 4.10.

	F1		F2		F3	
	Center Frequency	Bandwidth	Center Frequency	Bandwidth	Center Frequency	Bandwidth
/ə/ Low F3	500	80	1500	90	2040	150
/ə/ High F3	500	80	1500	90	3179	150
/o/ Low F3	497	80	938	90	2011	150
/o/ High F3	497	80	938	90	3118	150

Table 4.2: Experiment 2: Spectral Characteristics of the Vowel Tokens

The center frequency and bandwidth for each of the first three formants are provided in Hertz. The stimuli were synthesized using KLSyn (K. N. Stevens & Bickley, 1991), a user interface for the HLSyn speech synthesizer. The formant ratio calculations were performed in Mel frequency space and then converted back into Hertz for the speech synthesis.

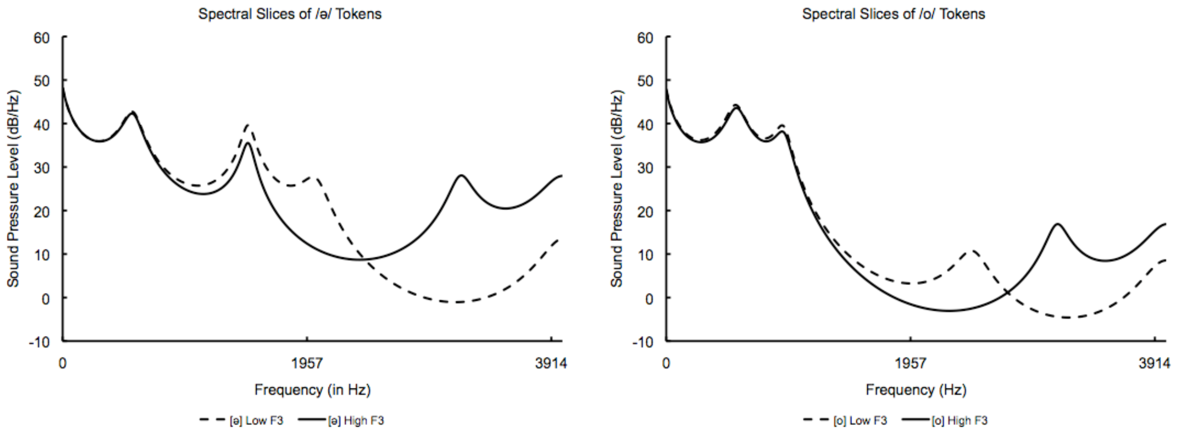


Figure 4.10: Spectral slices of the two /ə/ tokens and two /o/ tokens.

LPC-based spectral envelopes of the vowel sounds used in Experiment 2. The solid line indicates the token with a high F3 (smaller F1/F3 ratio) and the dashed line indicates the token with a low F3 (larger F1/F3) ratio. Spectral envelopes have been smoothed with a six pole LPC filter.

Procedure

The procedure was identical to Experiment 1.

Recording and Analysis

The recording parameters and analysis procedures used in Experiment 2 were identical to those used in Experiment 1. All trials with artifacts above 2.5 pT in the noise-reduced data were eliminated from analysis (5.2% of the total data). The filtering and baseline correct parameters are identical to those used in Experiment 1.

Results

Given the results from Experiment 1 and our hypothesis that the perceptual system displays a greater sensitivity to formant ratios for vowels located in more densely populated regions of the vowel space, we predict to find a reliable difference between the two tokens of /o/, with the token with a lower F3 (larger F1/F3) eliciting a shorter M100 latency, while I expect to replicate the null difference for the two tokens of /ə/ that we found in Experiment 1.

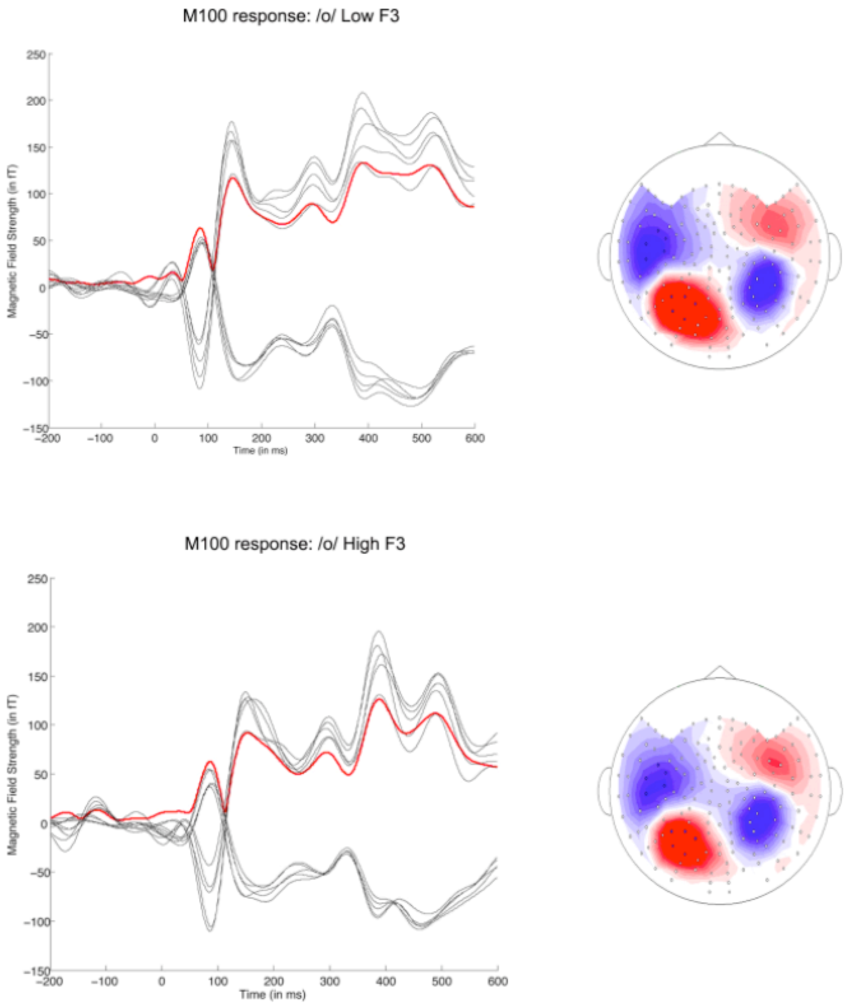


Figure 4.11: Evoked M100 Temporal Waveform and Magnetic Field Contour for Representative Subject.

Temporal waveform from ten left hemisphere channels (RMS solid red line superimposed) and the magnetic field distribution at peak latency of the M100. Waveform and field distribution to /o/ token with Low F3 (larger F1/F3 ratio) presented along the top row and waveform and field distribution to /o/ token with High F3 (smaller F1/F3 ratio) presented along the bottom row. For the magnetic field distributions, red indicates outgoing magnetic field source and blue indicated ingoing magnetic field sink.

Pooling across the channels from both hemispheres (20 channels), we find a reliable difference in the latency of the M100 between the two tokens of /o/ (One-tailed t-test; t(8) = 2.692, p < 0.02), where the token with a lower F3 (larger F1/F3 ratio; M = 132 ms) elicits a shorter latency than the token with a higher F3 (smaller F1/F3 ratio; M = 141 ms). Analyzing the hemispheres independently, we find a reliable difference between the two tokens of /o/ in both the left (t(8) = 2.929, $p < 0.01$) and right hemispheres (t(8) = 2.326, $p < 0.05$). Given the relatively small sample size, we also calculated the statistics using the non-parametric one-tailed Wilcoxon Signed-Rank test, and the parametric results are confirmed (Both hemispheres: V = 41.5, $p < 0.02$; LH: V = 40, $p < 0.02$; RH: V = 41.5; $p < 0.02$). We did not find a reliable difference in the amplitude of the RMS to the evoked M100 when comparing between the two tokens of /o/ (Both hemispheres: t(8) = -0.851, $p = 0.42$; LH: t(8) = -0.516, $p = 0.62$; RH: t(8) = -1.48, $p = 0.18$).

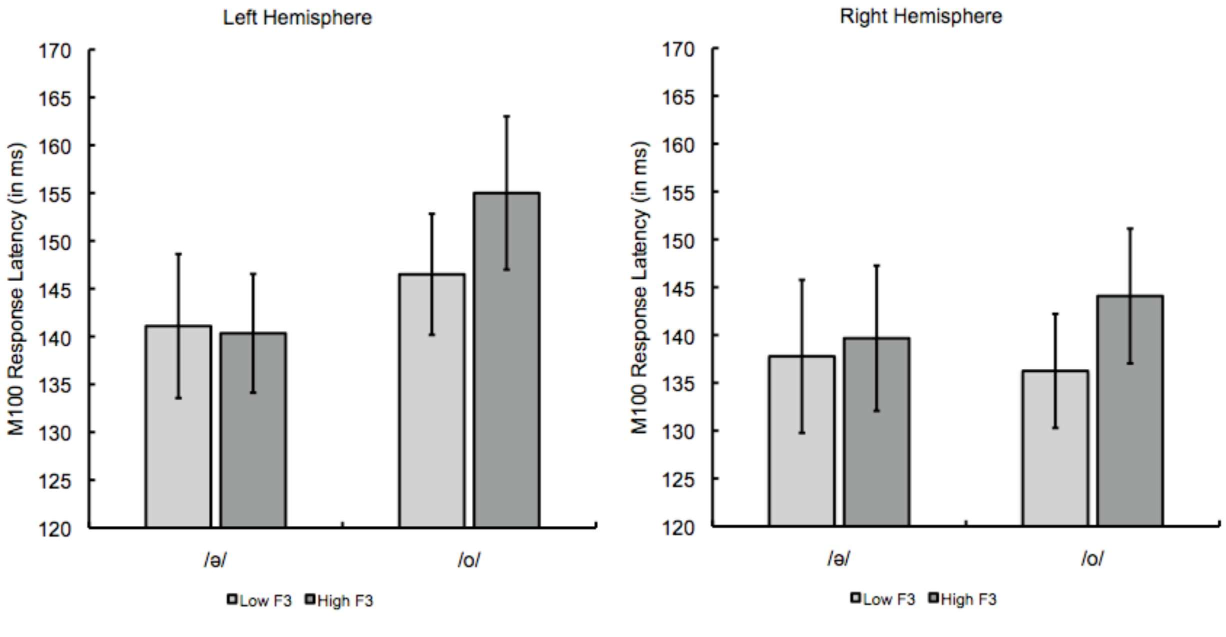


Figure 4.12: M100 Response Latencies by Vowel Type in Experiment 1.

Mean M100 response latencies across participants to the vowel tokens in Experiment 2. Light gray bars refer to tokens with a Low F3 (large F1/F3 ratio) and dark gray bars refer to tokens with a high F3 (small F1/F3 ratio). Error bars represent one standard error of mean.

I was able to replicate the differences we found for the two tokens of /ɛ/ in Experiment 1 with the two tokens of /o/ in the predicted direction: shorter M100 latencies for the tokens with a larger F1/F3 ratio. Moreover, we replicated the null effects for the two tokens of /ə/ in Experiment 1 again in Experiment 2. Again, pooling across both hemispheres, we fail to find a reliable difference between the two tokens of /ə/ (Low F3: M = 140 ms, High F3: M = 137 ms, t(8) = -1.065, *p* = 0.84), and we find no differences when looking at the hemispheres independently (LH: t(8) = -0.253; *p* = 0.6; RH: t(8) = -0.321, *p* = 0.62). Additionally, we found reliable differences in the amplitude of the RMS of the evoked M100 when comparing between the two tokens of /ə/ (Both hemispheres: t(8) = -1.944, *p* = 0.09; LH: t(8) = -0.916, *p* = 0.39; RH: t(8) = -1.217, *p* = 0.26). Therefore, we were able to extend the findings for /ɛ/ in Experiment 1 to /o/ in Experiment 2 and we replicated the null effect for /ə/ from Experiment 1 in Experiment 2. The findings from Experiment 2 confirm that 1) the auditory perceptual system (at least the neurobiological generators of the M100) is sensitive to formant ratios, and in particular at least the F1/F3 ratio, and 2) that the perceptual system is more sensitive to formant ratios in regions of the vowel space that are more densely populated.

M100 ECD Source Location

To assess whether the vowels presented to participants in Experiment 2 elicited differences in their source localization as well as the latency of the evoked M100, we calculated the ECD solution based on a spherical model of the head for the four distinct vowel tokens on an intra-subject and intra-hemispheric basis (minimum GoF =

90%, mean GoF = 95%; one participant was excluded as we were unable to calculate a fit with a GoF > 90%, statistics on source n = 8). Our only hypothesized prediction is that the tokens of the mid-vowel /ə/ might localize to a more anterior portion of the anterior-posterior plane than the tokens of the back vowel /o/. Performing a repeated measures ANOVA with the factors Vowel (/ə/ and /o/) and F3 (‘High’ and ‘Low’), in the left hemisphere, we find no main effect of Vowel (F(1,7) = 0.017, *p* = 0.9) or F3 (F(1,7) = 1.318, *p* = 0.29) and no interaction of Vowel × F3 (F(1,7) = 0.875, *p* = 0.38) along the lateral-medial plane (*x*-axis; sagittal).

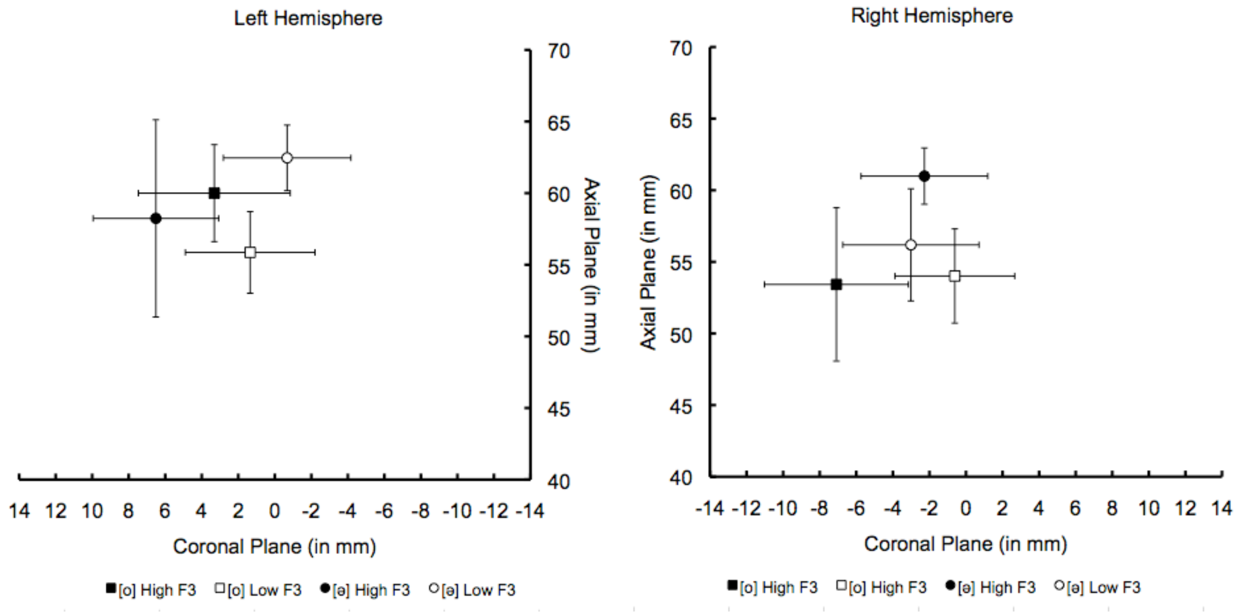


Figure 4.13: Mean ECD Locations: Coronal (y) Planes and Axial (z) Planes.

Plot of the mean ECD source locations for each vowel token in Experiment 2 plotted along the Coronal (y; anterior-posterior) and Axial (z; superior-inferior) planes. Error bars represent one standard error of the mean.

Along the superior-inferior dimension (*z*-axis; axial), we find no main effect of Vowel (F(1,7) = 1.376, *p* = 0.28) or F3 (F(1,7) = 0.0002, *p* = 0.99) and no interaction of Vowel × F3 (F(1,7) = 2.805, $p = 0.14$). Finally, along the anterior-posterior dimension (*y*-axis; coronal plane), where we might expect to find a difference given the results reported in Obleser, et al. (2004), we find a marginal effect of F3 (F(1,7) = 5.403, $p = 0.053$), which was unpredicted (we would have predicted an effect of vowel), no main effect of Vowel (F(1,7) = 0.026, $p = 0.88$) and no Vowel × F3 interaction (F(1,7) = 0.903, $p = 0.37$). To determine if there were any directional differences in the location of the ECD between the vowels along the anterior-posterior dimension, we performed a sign test on the different vowels types. We find a strong directional difference for the tokens with a High F3 (S = 1; $p < 0.05$), in the opposite direction, with the ECD of the token of /o/ with a High F3 localizing to a more anterior position along the anterior-posterior dimension than the token of / ə / with a High F3. We find no difference between the tokens with a Low F3 (S = 5; $p = 0.86$).

In the right hemisphere, along the lateral-medial (*x*-axis; axial) plane, we find no main effect of Vowel (F(1,7) = 0.036, $p = 0.85$) or F3 (F(1,7) = 0.682, $p = 0.44$) and no interaction of Vowel × F3 (F(1,7) = 1.083, $p = 0.33$). In the superior-inferior (*z*-axis; sagittal) plane in the right hemisphere, we again find a marginal main effect of Vowel (F(1,7) = 4.39, $p = 0.07$) but no main effect of F3 (F(1,7) = 0.37, $p = 0.56$) and no interaction of Vowel × F3 (F(1,7) = 0.439, $p = 0.53$). Finally, along the anterior-posterior (*y*-axis; coronal) plane, we find no main effect of Vowel (F(1,7) = 0.105, $p = 0.76$) or F3 (F(1,7) = 1.482, $p = 0.27$) and no interaction of Vowel × F3 (F(1,7) = 1.506, $p = 0.26$). In a sign test, to determine if there is a directional difference in the source of the ECD to vowels located along the anterior-posterior dimension, we found no effect between /ə/ and /o/ with Low F3s (S = 6; *p* = 0.14), nor did we find an effect between /ə/ and /o/ with High F3s (S = 4; *p* = 0.64).

Discussion

The motivation for Experiment 2 was to determine whether our hypothesis regarding the correlation of the density of categories in perceptual space and the sensitivity of the perceptual system to formant ratios was on the right track. Recall that in Experiment 1, we found a significant M100 latency difference for the /ɛ/ token with a larger F1/F3 ratio but not for the /ə/ token with a larger F1/F3 ratio. If an adequate explanation for the findings in Experiment 1 is that the sensitivity of our perceptual system to the F1/F3 ratio is a function of how dense the space is, and consequently, how much more competitive categorization is, then we also predict to find a significant difference for tokens of /o/ that vary on the F1/F3 ratio. As predicted, the token of /o/ with a larger F1/F3 ratio elicited a shorter M100 latency than the token of /o/ with a smaller F1/F3 ratio. And equally important, we replicated the null effect for /ə/. This reaffirms our findings from Experiment 1 that the auditory perceptual system is sensitive to F1/F3 ratios, lending further support to use of ratios in normalization algorithms. In particular, it demonstrates that formant ratios are psychologically plausible computations that can be exploited in the course of speaker normalization.

General Discussion

Speaker normalization has been a long-standing problem in speech perception research (Johnson, 2005). In particular, the perceptual and biological computations responsible for mapping highly variable acoustic input onto phonetic and phonological representations remain poorly understood (Sussman, 2000). Within the domain of vowel perception and normalization, a variety of different proposals have been offered to account for speaker dependent variation (Strange, 1989; Disner, 1980; Miller, 1989; Irino & Patterson, 2002; Rosner & Pickering, 1994; Nearey, 1989). In this chapter, we revisit a proposal that has been sporadically proposed in the literature: listeners are sensitive to the relative differences between formants (formant ratios) and not their absolute values (Miller, 1989; Peterson, 1951, 1961; Peterson & Barney, 1952; Syrdal & Gopal, 1986; Lloyd, 1890). A number of different formant ratio algorithms have been previously discussed in the literature. We propose a (relatively) novel formant ratio algorithm in which the first (F1) and second (F2) formants are ratioed against the third (F3). Higher formants, such as F3, may act as an adequate normalizing factor (Deng & O'Shaughnessy, 2003) and had been, at least impressionistically, judged to do a relatively good job at eliminating speaker dependent variation (Peterson, 1951), the kind of variation found in looking at acoustic differences between men, women and children. In order for a normalization algorithm to be adequate, we suggest that they must satisfy two criteria. First, it must adequately eliminate speaker dependent variation. Second, auditory cortex and the perceptual system must be able to perform the computation. Much of the previous work on formant ratios have been algorithms that require large corpora to adequately eliminate speaker variation (Miller, 1989). One of the intuitive advantages of the algorithm we propose here is that it appears to be able act a quick and dirty but efficient computation for online speaker normalization that can be performed with little exposure to a given speaker, which is consistent with what we know about dialect identification (Purnell, et al., 1999), the perceptual abilities of infants (Kuhl, 1979, 1983), and listeners' abilities to make speaker size estimates (Ives, et al., 2005; Smith, et al., 2005).

In this chapter, we investigated whether the perceptual system is sensitive to the F1/F3 ratio (the more novel of the two ratios; F2/F3 has appeared in previous ratio algorithms (Syrdal & Gopal, 1986; Miller, 1989)). We reported data from two MEG experiments that demonstrate that the neurobiological generators of the M100, an early, auditory evoked neuromagnetic component is sensitive to modulation of the F1/F3 ratio. The M100 had been previously reported to show sensitivity to the frequency of F1 in vowel perception (Govindarajan, et al., 1998; Roberts, et al., 2000; Poeppel, et al., 1997; Hannu Tiitinen, et al., 2005; Diesch, et al., 1996; Roberts, et al., 2004). Given our hypothesis regarding the algorithm involved in vowel normalization and the consequential representational nature of the vowel space (F1/F3 by F2/F3), we reinterpreted the previous MEG findings to conclude that the M100 is actually sensitive to the F1/F3 ratio and not F1 alone. The frequency of the third formant (F3) was not typically modulated in the previous MEG experiments, only F1. Therefore, we hypothesized that if we varied the value of the F3, and consequently, the F1/F3, we should be able to modulate the latency of the M100 in a predicted direction if the neurobiological generators of the M100 are sensitive to the F1/F3 ratio.

Our findings suggests the perceptual system can calculate formant ratios, lending further support to the notion that this is a plausible normalization algorithm, and moreover, that the M100 is sensitive to the F1/F3 ratio and not F1 alone. Furthermore, we calculated the statistical effectiveness of this algorithm in eliminating variance that is a function of the age and gender of a speaker on a large corpus of productions of American English vowels (Hillenbrand, et al., 1995). While the statistical analysis was perfunctory in many respects (e.g., we did not calculate how well the vowel space categorizes or how well particular tokens are classified as is normally done), which was beyond the scope of this chapter, the calculations suggest that large amounts of the speaker dependent variation when we compare vowel utterances across different talkers that as a function of age and gender was eliminated.

While we found that auditory cortex is sensitive to modulations of the F1/F3 ratio, the pattern of effects suggest a more nuanced conclusion. In the first experiment, we compared the M100 response latency to two tokens each of /ə/ and /ɛ/. The values for F1 and F2 were held constant while we moved F3 4% higher and lower in Mel space, such that the two tokens of a vowel type were 8% apart in F1/F3 ratio space. We found a reliable difference in the predicted direction only for the front-mid vowel /ɛ/, but no difference between in the response latency of the M100 for the two tokens of /ə/. As a result of this asymmetric result and the direction of the pattern, we hypothesized that the perceptual system displays heightened sensitivity to modulations of the F1/F3 ratio only when mapping acoustic information into more crowded regions of the vowel space. Experiment 2 was designed to test this hypothesis. We predicted that if the sensitivity to formant ratios is dependent upon how densely populated the region of vowel space the tokens are being mapped into, then we should expect to find a similar pattern of results when we compare tokens of a back vowel which vary on F3. Therefore, in Experiment 2, we tested two tokens of /o/, a mid-back vowel, that were 8% apart in Mel vowel space and attempted to replicate the null effect for the tokens of /ə/ with a new set of participants. Again, we find a reliable difference in the latency of the M100 between the two tokens of /o/ in the predicted direction and we replicated the null effect for /ə/. Our findings demonstrate that while the perceptual system is sensitive to formant ratios, its sensitivity is not equal across the space. Instead, the neurobiological generators of the M100 were sensitive only to vowel tokens whose acoustic structure mapped either onto the front or back of the space. Vowel tokens that mapped into the center of the space did not systematically modulate the latency of the M100. To place these findings within a theoretical framework, it so happens that in English, the front and back portions of the vowel space are more densely populated and therefore, categorization is more competitive. In other words, the acoustic distribution of a vowel can afford to be more diffuse in central portions of the vowel space where no other categories exist, as compared to more densely populated regions of the space, where more different vowel categories are located. This provides an intuitive explanation for why we might find a greater sensitivity of the neurobiological generators of the M100 to vowels located in the front and back of the vowel space as compared with vowels located in the center of the space.

An alternative explanation of the results we report would be that the M100 response latency is sensitive to all aspects of the spectrum, and therefore is also sensitive to modulations of F3 or perhaps even to differences in the power spectral density of the vowel tokens (see Roberts, et al., 2000 for results that suggest the M100 is sensitive to PSD). However, given that the differences in F3 between the tokens for each category (/ɛ/: Δ = 1091 Hz; /ə/: Δ = 1139 Hz; /o/: Δ = 1107 Hz) are roughly equivalent in raw Hz space and moreover, the differences between tokens for are all equal in Mel space (8% difference in the Mel space) this explanation will not adequately account for the M100 latency findings. Additionally, the differences in the central moment of the power spectral density of the tokens (PSD; /ɛ/: Δ = 11 Hz; /ə/: Δ = 14 Hz; /o/: Δ = 10 Hz) cannot account for the differences either, as /o/ has a smaller difference than /ə/ and yet, we found a reliable difference in the M100 response latency for /o/ and not for /ə/. While we accept that the overall power spectral density may help drive the response (differences in formant ratios will lead to differences in power spectral densities), we argue that our results suggest that it cannot be the entire answer.

We also wish to be clear that we are not proposing that this is the only algorithm or the sole computation in which listeners employ to normalize across vowel tokens spoken by different individuals. For example, the fundamental frequency, or f0 does not play a role in the current schema, while it has been present in many of the previous accounts of vowel normalization (Miller, 1989)

As a point about the M100 component itself, we can be confident that the M100 is not sensitive only to F1, but that higher regions of the frequency space also play a role in modulating its latency. In particular, we would conclude that the response latency of the M100 indexes more abstract computations that have been performed on the stimulus and in fact reflect complex representational schemas in auditory cortex. This conclusion is consistent with other work done on the relation between the M100 and F1 (Roberts, et al., 2004) and findings that demonstrate that the M100 is sensitive to differences in the inferred pitch of complex stimuli that are missing a fundamental component (Monahan, et al., 2008; Fujioka, et al., 2003).

The problem of how (and where, to an extent) the brain computes formant ratios appears, on the surface at least, to be a tractable one. The formant ratio solution seems to provide a point where biology and psycholinguistics could be fruitfully combined to provide a fairly complete account of some perceptual linguistic phenomenon. The brain is able to extract frequency information from the incoming auditory stimulus (Giard, et al., 1995). And presumably, the brain performs computations over these extracted frequency components. In the case of formant ratios, the brain must extract the frequency information of the formants (which we know is done, because we do have the perceptual experience of hearing different vowels, which differ only on formant information), possibly tag the different formants and compute the difference between F1 and F3 and F2 and F3. Subtraction (if the spectral peaks are transformed into log space) does not seem like an impossible computation for the brain to perform. Moreover, the fact that we find effects of these ratio differences by the M100 suggests that this computation is done by the time the auditory information reaches the M100 generators in auditory cortex. Essentially, then, we should seek out neural circuits that have the ability to compute differences between frequency components, and these neural circuits should be located early in the auditory processing stream. Whether or not we can find a cortical map that is representationally sympathetic to a perceptual map ((Eulitz, et al., 1995)) is up for debate. We should, however, be able to find neural circuits that can perform this algorithm that we have proposed in this chapter.

Conclusion

In short, we set out to approach the problem of speaker normalization from an online speech perception perspective. In other words, we were interested in testing whether the auditory perceptual system, and in particular the neurobiological generators of the M100 located in auditory cortex, were sensitive to formant ratios (specifically F1/F3). We found that the auditory perceptual system, however, shows differential sensitivity to formant ratios depending upon where in the space the vowels are located. In particular, we found significant M100 latency differences to /ɛ/ and /o/ but not to /ə/. Further research is required to find evidence for the F2/F3 ratio (we are unaware of a dependent measure in the electrophysiological literature sensitive to F2). While we are hesitant to conclude that this is the algorithm wholly responsible for successfully eliminating variance based on inter-speaker variation in vowel perception, we suggest that the exploitation of higher formants, in particular F3, in vowel normalization could provide valuable insight into furthering our understanding of speaker normalization.

Chapter 5: Early Detection of Phonological Violations*

Introduction

Languages place constraints on acceptable and unacceptable sequences of sounds. For example, in English, [tr] is an acceptable consonant sequence at the beginning of a word, whereas [rt] is not; at the ends of words, the situation is reversed, [rt] is acceptable, but [tr] is not. Some of these constraints are specific to individual languages, while others are universally attested, that is, they seem to occur in every extant human language. Many studies, both behavioral and neurophysiological, have demonstrated the role of language-specific phonetic inventories in speech perception (Werker & Tees, 1984; Hacquard, et al., 2007; Kuhl, 2004; Kuhl, et al., 1992; Kazanina, et al., 2006; Näätänen, et al., 1997), but relatively few have discussed the role and time-course that the knowledge of phonological constraints play in the auditory mapping between acoustic input and linguistic representations. In this chapter, we present magnetoencephalographic (MEG) data that suggests that auditory cortex is sensitive to violations of phonological constraints as early as 150 ms post-onset of the violating segment. We take this evidence to suggest that listeners generate predictions about relatively abstract properties of the speech signal based on higher-order knowledge

* Submitted as:

Monahan, P.J., S.-O. Hwang & W.J. Idardi (submitted). Predicting Speech: Neural Correlates of Voicing Mismatch using MEG. *Brain Research*.

(i.e., knowledge of constraints on sound sequences), and this knowledge constrains early auditory cortical processing of speech sounds.

Several earlier behavioral studies have demonstrated that psychophysical measures index such abstract phonological knowledge and have suggested that listeners are sensitive to violations of language-specific phonological constraints (Lahiri & Marslen-Wilson, 1991; Fowler & Brown, 2000; Gow, 2001; Gaskell & Snoeren, 2008). A few of these experiments have exploited the anticipatory nasalization rule in English, which accounts for the fact that typically oral vowels are produced with additional air-flow through the nasal cavity, and hence, become “nasalized vowels”, when they precede nasal consonants (i.e., [m n ŋ]). One question that could be asked regarding these structures is whether listeners can use the information on the vowel (namely, that it is nasalized) to predict that the upcoming segment is a nasal consonant in real-time speech perception. Behaviorally, it has been demonstrated using a consonant identification task that native speakers of English encounter most difficulty when they perceive a nasal consonant following an oral vowel (Fowler & Brown, 2000), and moreover, the presence of nasalization on the vowel leads to more nasal responses in a gating task prior to the perception of the critical consonant (i.e., the nasal consonant) (Lahiri & Marslen-Wilson, 1991). These findings suggest that listeners, to some extent, are able to use this language specific phonetic information to predict the nature of the upcoming speech signal.

More recently, electrophysiological techniques (MEG/EEG) have been employed to ask questions about the neural time-course of these violations of phonological expectation (Flagg, Cardy, & Roberts, 2005; Mitterer & Blomert, 2003;

Tavabi, Elling, Dobel, Pantev, & Zwitserlood, 2009). Flagg, et al. (2005) exploited the pattern described above where pre-nasal vowels are nasalized in English and measured the electrophysiological latencies of the response peaks in MEG to congruent (i.e., [aba], [ãma]) and incongruent (i.e., [ãba], [ama]) VCV sequences. Overall, latencies in the time-window of 50-100 ms post-onset of the consonant were shorter for the congruent as opposed to incongruent sound sequences. In a different experiment, Mitterer & Blomert (2003) found that violations of expected nasal place assimilation patterns between words elicited a larger MMN amplitude (approximately 100-200 ms after onset of the consonant) than adherence to assimilation patterns in Dutch listeners. These previous electrophysiology results (Flagg, et al., 2005; Mitterer & Blomert, 2003) suggest that although cortical responses that reflect phonological processing are early, the types of responses measured (latency versus amplitude) and the time-window in which differences were found varied. To date, the precise nature of the time course of the role of phonological knowledge in speech perception remains poorly understood.

In the present study, we use a pervasive cross-linguistic generalization as our case study: syllable-final obstruent consonant clusters must agree in their specification for voicing (i.e., whether or not the vocal folds vibrate during the production of the speech sound). That is, in all known languages, word-final obstruent consonant phonetic sequences that agree in voicing, such as [dz] (both voiced) and [ts] (both voiceless), are acceptable, while those that disagree are not: *[ds] ([d]: voiced; [s]: voiceless) and *[tz] ([t]: voiceless; [z]: voiced). Traditional linguistic analyses have taken this constraint to be the product of a phonological rule of assimilation. Typically, the feature [-voice] on voiceless consonants spreads to the word final consonant (Chomsky & Halle, 1968).

In a previous behavioral study (Hwang, S.O., Monahan, P.J., Idsardi, W.J., submitted), participants were presented with congruent (e.g., [udz], [uts]) and incongruent (e.g., [uds], [utz]) tokens and asked to respond whether they perceived [z] or [s]. In both the reaction time and accuracy data, we found that English listeners are sensitive to these violations, but only when the first consonant of the sequence is voiced (e.g., [d]). That is, we found an asymmetric pattern of results: responses to the congruent pair were inconsistent ([udz] was faster and more accurate than [uts]), and more interestingly, the incongruent pair also did not behave alike ([uds] was slower and less accurate than [utz]). Moreover, there was no difference between the grammatically acceptable [uts] and the grammatically unacceptable [utz]. We suggest that this particular pattern of results is anticipated if one assumes that phonological representations can be underspecified. That is, predictable distinctive features (e.g., [-voice], [coronal], etc.) are not a part of a sound's representation in long-term memory (Archangeli, 1988; Lahiri, 2007; Lahiri & Reetz, 2002). To illustrate, because the nasal consonant /n/ often undergoes place assimilation (/in+kʌmplit/ ⟶ [/iŋkʌmplit] 'incomplete') with the following sound (i.e., it takes on the place of articulation of the following consonant), it has been hypothesized that [coronal] is not specified for /n/. Therefore, the long-term memory representation for the phonological segment /n/ does not contain a specification for where in the mouth it is articulated because its place is predictable provided a phonological context. Only the features that are unpredictable count toward a particular sound's long-term representation. Returning to our previous behavioral results, the particular pattern is predicted if we combine the proposals that the feature [voice] is specified for only voiced consonants (e.g., [d], [z]) (Lombardi, 1995; Mester & Itô, 1989) with the hypothesis that only represented features can be used as the basis to make predictions about the upcoming speech signal (Lahiri & Reetz, 2002). More concretely, the feature [voice] on /d/ can be exploited to predict that the following sound is going to be voiced (i.e., [z]), consistent with the universal constraint. The prediction is met when the listener encounters /z/ and is violated when the listener encounters /s/. On the other hand, when a listener encounters /t/, where there is no feature for voicing in the long-term representation, and therefore, no prediction can be generated regarding whether the following sound is going to be voiced or voiceless. Asymmetric results of this sort are not uncommon, even using electrophysiological techniques, and are typically taken to support underspecified long-term representations (Eulitz & Lahiri, 2004; Friedrich, Eulitz, & Lahiri, 2006).

Using MEG, we examine the time course of the role of phonological knowledge in speech perception for this specific and putatively universal constraint (syllable final obstruent clusters must agree in voicing) and aim to better understand whether neurophysiological measures reflect a sensitivity to constraints on how speech sounds sequence in early auditory perception. Such findings contribute to a better understanding of the time course and nature of the mechanisms involved in mapping acoustic signals onto linguistic representations (Poeppel & Monahan, 2008; Obleser & Eisner, 2009; Poeppel, Idsardi, & van Wassenhove, 2008; Hickok & Poeppel, 2007).

Experimental Procedure

Participants

Fourteen monolingual native speakers of American English participated in this study (11 female; mean age: 20.8 years) and were included in the analysis. An additional six participants took part but were excluded from the analysis (3 for poor magnetic field contours; 2 for poor M100 responses to the vowel; 1 for poor M100 to the auditory localizer pretest). These are metrics that we use to assure that the participant is providing a reliable auditory response to the speech tokens, specifically, and auditory tokens, generally. Each participant tested strongly right handed on the Edinburgh Handedness Survey (Oldfield, 1971) and provided written informed consent. Subjects received either course credit or $10 for their participation.

Materials

A male native speaker of American English recorded natural utterances of the English non-words [ups], [uts], [uks] and [ubz], [udz], [ugz]. These recordings were edited with Praat (Boersma, 2001) to create tokens with voicing agreement (e.g., [uts], [udz]) and voicing disagreement (e.g., [utz], [uds]). The tokens with voicing agreement were spliced together from two tokens of the same type (e.g., [ut] was spliced from [uts] and combined with [s] which was spliced from a different token of [uts]). The tokens with voicing disagreement were cross-spliced together from two natural tokens (e.g., [ut] was spliced from [uts] and combined with [z] which was spliced from [udz]). This was done for each place of articulation. There were a total of twelve items in this study. In the end, all sounds were edited to eliminate any response bias toward edited speech stimuli. Additionally, the stimuli were edited such that each segment was 100 ms in duration (300 ms total for each VCC syllable) to eliminate any response bias based on stimulus length alone. The stop bursts were removed and each token was gradually ramped so that the vowel had a 20 ms fade-in and the final fricative had a 20 ms fade-out.

	Attested		Unattested	
Condition	UTS	UDZ	UTZ	UDS
Labial	[ups]	[ubz]	[upz]	[ubs]
Coronal	[uts]	[udz]	[utz]	[uds]
Dorsal	[uks]	[ugz]	[ukz]	[ugs]

Table 5.1: List of conditions in the experiment by place of articulation

A male native speaker of American English recorded natural utterances of the acceptable English non-words. Each segment was 100 ms in duration so that the total duration was 300 ms. The final consonant of each recorded token was removed and spliced together to create tokens with voicing agreement (attested) and voicing disagreement (unattested).

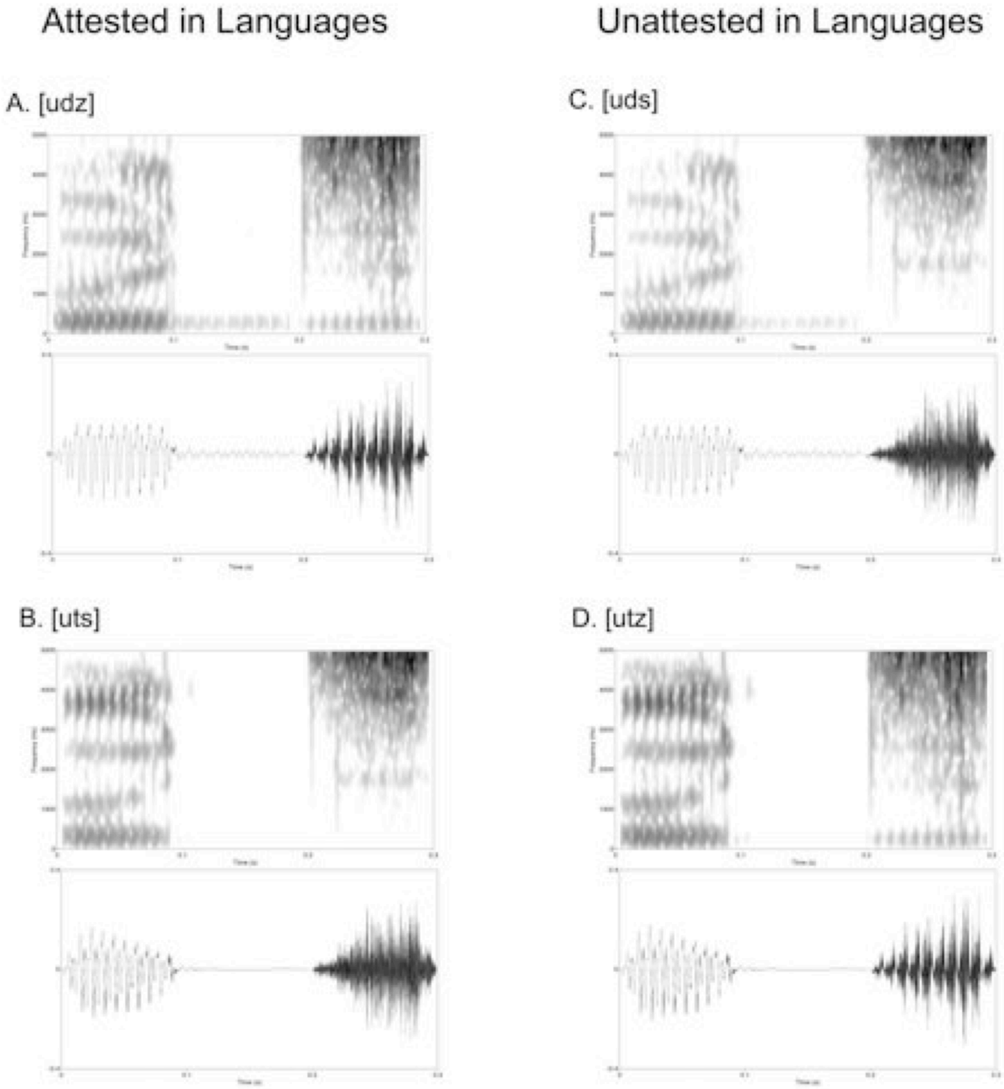


Figure 5.1: Spectrograms and waveforms for four conditions in the experiment.

The conditions on the left (A and B) are cross-linguistically attested (the two consonants agree in voicing). These tokens were created by splicing (A) [ud] with [z] from two tokens of [udz] and (B) [ut] with [s] from two tokens of [uts]. The conditions on the right (C and D) are unattested in the world's languages (the two consonants disagree in voicing). These tokens were created by cross-splicing (C) [ud] with [s] from a token of [udz] and a token of [uts], respectively and (D) [ut] with [z] from two tokens of [uts] and [udz] respectively. The amount of voicing is evident by the low frequency energy in the spectrograms and the periodicity in the waveforms between 0.1 and 0.2 sec in (A) and (C) and the respective absence in (B) and (D).

Procedure

Magnetoencephalographic recordings were acquired using a 157-channel whole-head axial gradiometer MEG system (Kanazawa Institute of Technology, Kanazawa, Japan). Participants lay supine in a dimly lit magnetically shielded room. Auditory stimuli were delivered binaurally via Etymotic ER3A insert earphones. Earphones were calibrated to have a flat frequency response between 50 Hz and 3100 Hz within the shielded room. Participants were first presented with an auditory localizer pretest to ensure adequate positioning of the head within the scanner. The pretest involved passive listening to a series of 250 Hz and 1 KHz sinusoids.

For the experiment, stimulus presentation included 150 randomized trials of each of the twelve tokens using Presentation® (Neurobehavioral Systems, Inc.) software. The inter-stimulus interval (ISI) varied pseudo-randomly between 450 ms and 1450 ms. Participants were asked to respond to a distracter 1 KHz sinusoid tone and listen passively to the speech stimuli. The neuromagnetic signal was sampled at 500 Hz with an online 200 Hz LPF and 60 Hz notch filter. Offline, the data were noise reduced using a multi-shift PCA noise reduction algorithm (de Cheveigné & Simon, 2007) and was band-pass filtered by a Hamming-window digital filter with frequency cut-offs at 0.03 Hz and 30 Hz. Stimulus-related epochs of 1100 ms (500 ms pre-trigger/fricative (i.e., [s z])) were averaged according to stimulus type. A portion of the epoched window (-400 to -300 ms) where no stimulus was present (-200 ms was onset of vowel) was used to baseline correct the averaged file. We were interested in whether the presence of a voiced obstruent (e.g., [d]) caused difficulty in the processing of a voiceless fricative (e.g., [s]) and vice versa. Therefore, the trigger was placed at the onset of the fricative (200 ms after the start of the token). Ten left hemisphere channels that best correlated with the neuromagnetic signal (five channels from the source and five channels from the sink of the dipole) were selected for statistical analysis based on the auditory pretest. The channels were selected on a participant-by-participant basis, but the selected channels were the same for all conditions for each individual subject. The RMS of the ten left hemisphere channels was calculated for each subject for each condition. Subsequently, the RMS temporal waveforms were averaged across participants for each condition, and a paired t-test was performed on the RMS of the planned comparisons for particular time windows.

Results

Given that we were interested in the response to the final fricative, and to eliminate the possibility that any differences found in the RMS of the MEG temporal waveform could be attributed to low-level acoustic properties of the stimulus, we compared only items that had the same final consonant (e.g., [uts] with [uds]; [udz] with [utz]). For each participant in each condition, we calculated the root mean square (RMS) amplitude of the MEG temporal waveforms from ten left hemisphere channels, selected on the basis of an auditory localizer pretest.

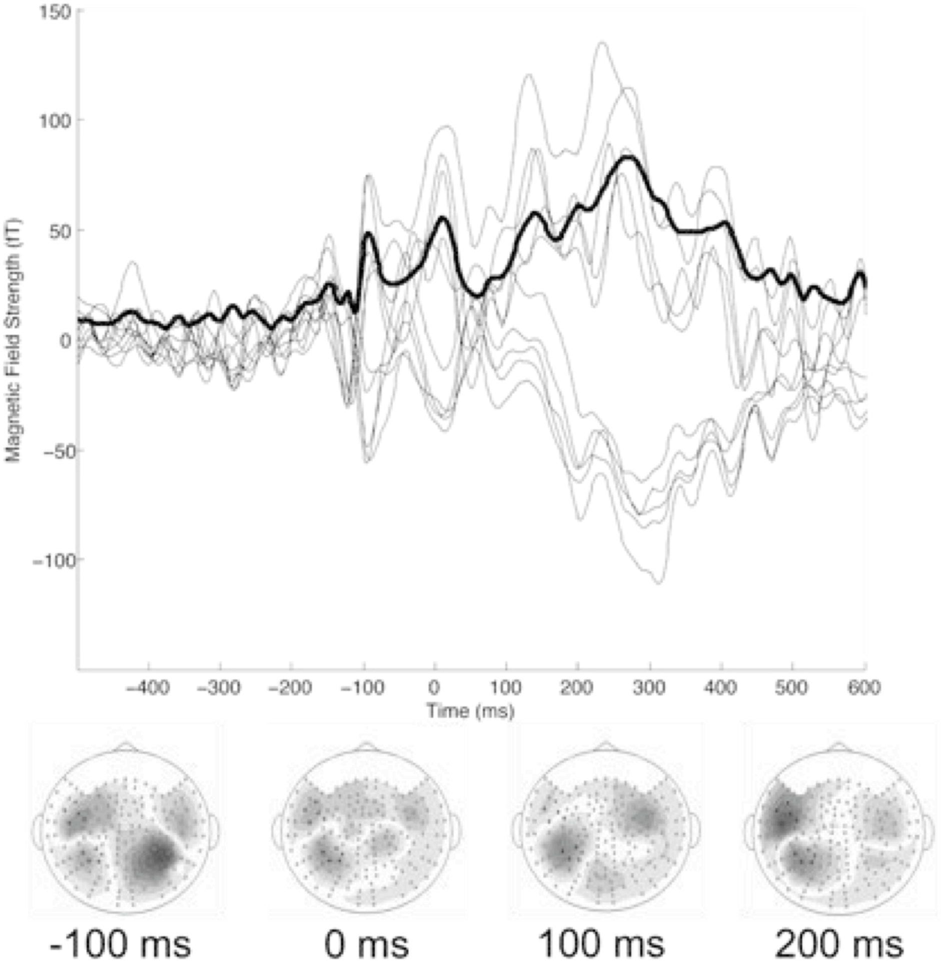


Figure 5.2: MEG waveform from a representative subject for one condition.

The waveform plot is an overlay of ten left-hemisphere sensors selected on the basis of an auditory localizer pretest. Bold line is the root mean square (RMS) of the ten channels. Magnetic field contour plots from an individual subject for four different time points are presented below the waveform plot (-100 ms: 100 ms post-onset of the vowel; 0 ms: 100 ms post-onset of first consonant; 100 ms: 100 ms post-onset of the second consonant; 200 ms: 200 ms post-onset of the second consonant). Dark gray refers to sink of magnetic field, while light gray refers to the source of the magnetic field.

Subsequently, we calculated the grand average RMS across participants (n=14) for each condition. For the conditions of interest, paired two-tailed t-tests were performed on the grand averaged RMS waveforms for specific time-windows. Two factors led to our choice of the 150-400 ms time window as the initial window of comparison. First, visual inspection of the grand averaged RMS waveforms suggested a large difference in this time window. Second, early auditory responses are known to occur before 150 ms (Roberts, et al., 2000) and higher-order lexical-semantic effects are traditionally seen by 350 to 400 ms (Lau, Phillips, & Poeppel, 2008). We hypothesize that whatever effects we might find would likely be due to mechanisms that operate between initial auditory processing and lexical access. Therefore, the time window of 150–400 ms provides the temporal boundary conditions within which we were looking for a systematic response modulation of the MEG temporal signal.

Collapsing across place of articulation of the stop consonant, we found a significant difference between UTS (UTS: [uts],[ups],[uks]; C1: unvoiced, C2: voiced) and UDS (UTS: [uds],[ubs],[ugs]; C1: voiced, C2: unvoiced) in the time-window of 150-400 ms post onset of the fricative (t = -2.68; $p < 0.02$).

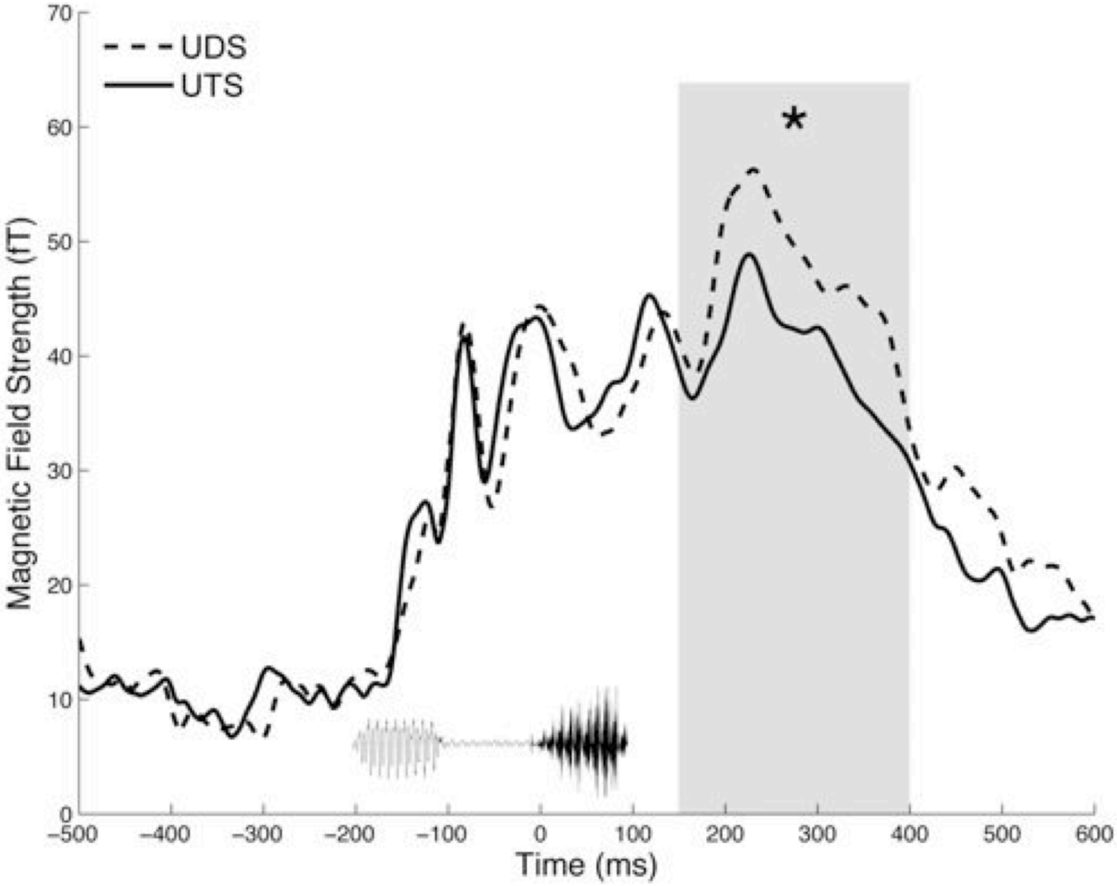


Figure 5.3: Overlay of the grand average MEG RMS temporal waveforms.

Conditions UDS (dotted line) and UTS (solid line) collapsed across place of articulation. Shaded area on the waveform plot designates region of significant difference between the two conditions. Acoustic waveform overlay on temporal waveform plot denotes stimulus presentation relative to neuromagnetic signal.

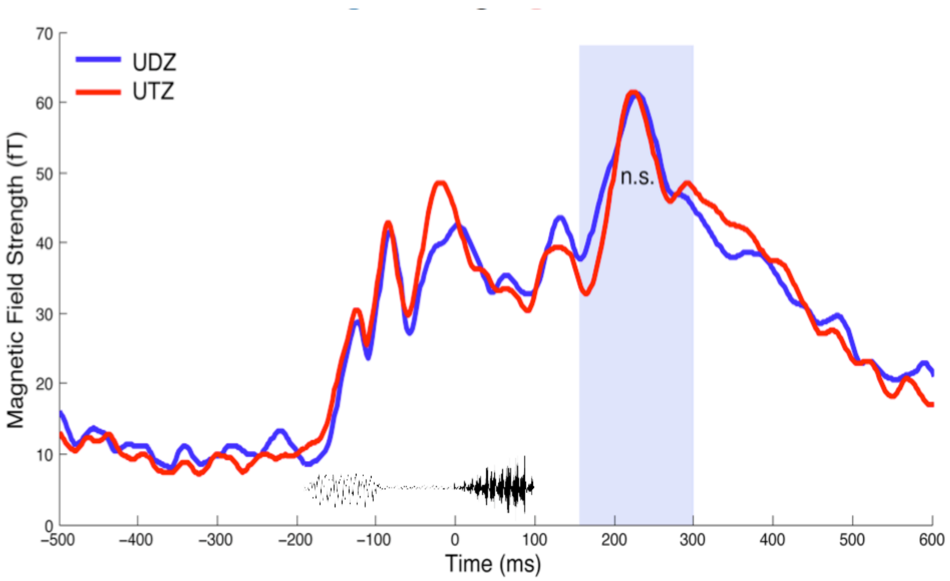


Figure 5.4: Overlay of the grand average MEG RMS temporal waveforms.

Conditions UDZ (blue line) and UTZ (red line) collapsed across place of articulation. Acoustic waveform overlay on temporal waveform plot denotes stimulus presentation relative to neuromagnetic signal.

For the stimuli in which the final consonant was voiceless (UTS/UDS), we found a significant effect for coronals (t=-3.01; p < 0.02), a marginal effect for labials (t=-1.93; p = 0.08) and no effect between the dorsal tokens (t=-0.95; p = 0.36) in the time window of 150-400 ms. In order to more precisely determine the time course of these effects, we tested smaller time windows by dividing the original time window in half: 150-275 ms and 275-400 ms. For the earlier time window (150-275 ms), we found a significant difference when collapsing across the place of articulation (t=-2.29; p < 0.05). Analyzing the individual places of articulation independently, we found marginal effects for the labial (t=-1.77; p = 0.10), coronal (t=-1.75; p = 0.10) and dorsal (t=-1.93; p = 0.08) pairs. In the later time window (275-400 ms), we again found a difference when collapsing across the place of articulation (t: -2.64; p < 0.05). Moreover, we found a reliable effect for the coronal pair (t=-3.20; p < 0.01) and no effect for the labial (t=-1.69; p = 0.12) and dorsal pairs (t=-0.15; p = 0.88). We did not, however, find a reliable difference between UTZ and UDZ in the larger time window (150-400 ms: t=0.27; *p* = 0.79) or in either of the smaller time windows (150-275 ms: t=-0.62; *p* =0.54; 275-400 ms: t=1.33; *p* = 0.21). No other differences were found in any other time-window.

Discussion

In the current experiment, we tested a cross-linguistically attested phonological constraint that requires syllable final consonant clusters to agree in voicing. To the best of our knowledge, this process has yet to be exploited in the understanding of the neural mechanisms underlying phonological processes. Analyzing the grand-averaged RMS of the MEG temporal waveforms, we found a reliable difference between congruent (i.e., UTS) and incongruent (i.e., UDS) syllables as early as 150 ms post-onset of the violating segment, in this case, the fricative (i.e., [s]). We take the differences between UTS and UDS to suggest that listeners can exploit their knowledge of phonological processes and representation to constrain early perceptual parses of the sensory input, as well as using this detailed knowledge to serve as the basis for generating hypotheses and predictions about the nature of the upcoming speech signal.

Consistent with an underspecification view of perception, we hypothesize that only features that are specified in the long-term representation for phonological segments can form the basis for the generation of predictions regarding the nature of the upcoming speech stimulus (Lahiri, 2007; Lahiri & Reetz, 2002). In this particular case, we adopt the proposal that [-voice] is not represented for voiceless consonants while [+voice] is represented for voiced consonants (Mester & Itô, 1989; Lombardi, 1995). Therefore, only sounds that contain [voice] (i.e., voiced consonants) can be used to predict the phonological quality of the upcoming sound, in particular, that the next consonant in the syllable must also be specified for [voice]. Sounds that do not have a specification for [voice] (i.e., voiceless consonants) cannot be used to predict whether the next consonant in the syllable is also voiceless. Consequently, from our data, it appears that when a listener unexpectedly encounters a voiceless consonant followed a consonant specified for [voice], they are surprised, and this violation of expectation is indexed by early cortical processing. When a listener encounters a sound that is not specified for [voice], no predictions can be made, and thus, we found no difference between UTZ and UTS. Moreover, it appears that this knowledge of phonological representations and constraints which act as the basis for these online predictions are reflected in early cortical processes.

Unlike our previous behavioral study (Hwang, S.O., Monahan, P.J., Idsardi, W.J., submitted), where two effects were found: processing facilitation for UDZ and processing difficulty for UDS (while UTS and UTZ showed no differences), the present MEG study only showed a difference between UDS in comparison with UTS. While we did not predict a lack of difference between UTZ and UDZ (it should be noted that the magnitude of the effect size in both the reaction time and accuracy data was significantly smaller for UDZ versus UTZ as opposed to UTS versus UDS in the previous behavioral experiment), this asymmetric result does, however, allow us to eliminate a possible alternative explanation for our findings in which the phonetic quality of the obstruent consonant (i.e., D versus T) is driving the differences in the

RMS of the MEG temporal waveform. If this alternative explanation were true, then we would have expected to also find a difference between UTZ and UDZ. We can thus be more confident that the difference we did find is attributable to a violation of expectation and not low-level acoustic properties of the stimulus.

The time course of our effects are consistent with the previous electrophysiological results (Flagg, et al., 2005; Mitterer & Blomert, 2003) that investigated the role of phonological knowledge in speech perception. Flagg et al. (2005) tested violations of the language specific constraint that pre-nasal vowels are nasalized in English. They found a reliable difference in the latency of the M50 in the MEG waveform to the consonant for the tokens [aba] compared with [ãba]. They failed to find a difference when the consonant was the nasal [m], however (i.e., [ãma] compared with [ama]). Given the complex nature of the evoked magnetic waveform to these stimuli, we are less confident that the component they identified was, in fact, the M50. The time-course of these effects, however, suggests that differences are evident in early cortical processing. Mitterer & Blomert (2003) used an MMN paradigm with Dutch speakers and found that unviable phonological assimilations (/n/ becoming [m] before /s/) elicited a mismatch negativity in a passive oddball paradigm, while viable phonological assimilations (/n/ becoming [m] before /b/) did not. The effects in Mitterer & Blomert (2003) were seen in the traditional MMN time window (~250 ms post-onset), again, suggesting that early cortical processes are sensitive to violations of a phonological constraint. Unlike the Flagg et al. (2005) results, the phonological process used by Mitterer & Blomert (2003) is generally cross-linguistically attested. That is, coronal nasal consonants (i.e., /n/) usually undergo assimilation to the place of articulation of a neighboring consonant, though the directionality of assimilation often differs on a language-by-language basis. The cross-linguistic nature of this assimilation pattern is evident by the fact that they also tested German listeners on the Dutch contrast and found similar results.

Conclusion

Convergent with other behavioral findings, these MEG results suggest that listeners make use of their knowledge of phonological constraints regarding sound sequences to predict the phonetic quality of the upcoming sound. Moreover, violations of these expectations are seen in early auditory cortical processes, as indexed by the RMS of the MEG temporal waveform, in particular, by 150 ms post-onset of the violating segment. These results provide further support to the idea that listeners actively utilize their phonological knowledge to parse the speech signal and make predictions regarding the phonetic quality of the upcoming segment.

Chapter 6: Conclusion

Understanding the cortical mechanisms underlying spoken word recognition has been a long-standing issue in cognitive science and (cognitive) neuroscience (Geshwind, 1970; R. H. Fitch, et al., 1997; Hickok & Poeppel, 2000, 2007; Näätänen, 2001; Scott & Johnsrude, 2003; Davis & Johnsrude, 2007; Poeppel, et al., 2008; Poeppel & Monahan, 2008; Obleser & Eisner, 2009). Electrophysiology (EEG/MEG) has proven to be an extremely powerful tool in assessing the processes and representations involved in mapping the time-varying acoustic waveform onto stored memory-representations sub-serving the cognitive and neurophysiological processes underlying speech perception (Phillips, 2001; Näätänen, 2001; Näätänen, et al., 2007). In this dissertation, I presented data demonstrating that even at the earliest (non-invasive) recordable stages of auditory processing, we can find evidence that auditory cortex is calculating abstract representations from the auditory signal. While the results from the experiments surely do not point to phonological abstractions (Chapter 5 notwithstanding), they do point toward the types of representations and computations that may lead to phonological abstraction. For example, in Chapter 4, I proposed a novel formant ratio algorithm intended to help solve the vowel normalization problem. In addition to providing a potential solution to this perceptual problem, it also proposed a novel representational space for the organization of vowel categories. Being able to tap into these early, automatic processes, we can begin to understand the earliest stages of auditory cortical computation. Moreover, the experiments proposed in Chapters 3 and 4 provided evidence that the even at the earliest stages of cortical processing (the M100 localizes to Heschl's Gyrus, planum temporale), we find evidence of auditory cortex performing abstract computations over the physical signal: abstraction at primary auditory cortex.

In Chapter 2, I briefly reviewed the literature on episodic models of speech perception, as well as recent perceptual and functional imaging data, which suggests that the strongest episodicist position is untenable. There must be prelexical abstraction (note that this does not preclude the encoding of fine acoustic structure; it is just adjunctive to the core abstract representation). Subsequently, I reviewed the electrophysiological literature (primarily the MMN), which also demonstrates cortical sensitivity to phonological representations. In Chapter 3, I presented neuromagnetic data that suggests that the response properties of the M100 are sensitive to the missing fundamental component in auditory perception. I found that the M100 latencies for the tone complexes (with a missing fundamental) matched their pure sinusoid counterparts, while also replicating the M100 temporal latency response curve found in previous studies. I understood these findings to suggest that listeners are reconstructing the inferred pitch by 100 ms after stimulus onset. In Chapter 4, I presented a novel solution to the vowel normalization problem and presented MEG data suggesting that auditory cortex is sensitive to the computations required by this algorithm (F3 is the normalizing factor; vowel space = F1/F3 by F2/F3). These findings also demonstrated that the perceptual system shows heightened sensitivity to tokens located in more densely populated regions of the vowel space. More work is certainly needed to better understand the perceptual and cognitive factors causing this effect. To the best of my knowledge, little is known about the relative sensitivity to different regions of the vowel space based on category distributions. Finally, in Chapter 5, I reported MEG results suggesting that early auditory cortical processing (~ 150 ms post-onset of the violating speech segment) is sensitive to a cross-linguistic constraint on sound sequencing (i.e., word final obstruent consonants must agree in their specification for [voice]). I argued that these findings suggest that listeners make highly specific, knowledge-based predictions about rather abstract anticipated properties of the upcoming speech signal and violations of these predictions are evident in early cortical processing.

In the case of the inferential pitch Chapter (3), it was clear that by 100 ms post-onset of a stimulus, a representational object of the physical stimulus had been constructed which included the missing fundamental component, given that the response properties of the M100 to the tone complexes missing the fundamental component tracked the response to the corresponding pure sinusoids, whose frequencies were at the inferred frequency of the tone complexes. In the vowel normalization experiment, early auditory cortex, as measured in the MEG temporal waveform, indexed a computation (ratio) performed over spectral peaks within the stimulus and no longer reflecting the physical properties of the spectral peaks. Finally, in Chapter 5, knowledge of phonological sound sequencing constraints had a distinct influence on the early cortical processing of complex syllables. In conclusion, the experiments presented in this dissertation have demonstrated that at the earliest stages of auditory cortical processing, we can find evidence of abstraction (the calculation of representational objects that are no longer faithful to the physical properties of the stimulus), which may ultimately allow us to better understand the nature of higher-order linguistic representations.

Bibliography

Adank, P., Smits, R., & van Hout, R. (2004). A Comparison of Vowel Normalization Procedures for Language Variation Research. *Journal of the Acoustical Society of America, 116*(5), 3099-3107.

Alain, C., Cortese, F., & Picton, T. W. (1998). Event-Related Brain Activity Associated with Auditory Pattern Processing. *Neuroreport, 26*(9), 3537-3541.

Alho, K., Sainio, K., Sajaniemi, N., Reinikainen, K., & Näätänen, R. (1990). Event-Related Brain Potential of Human Newborns to Pitch Change of an Acoustic Stimulus. *Electroencephalography and Clinical Neurophysiology, 77*(2), 151-155.

Alho, K., Winkler, I., Escera, C., Huotilainen, M., Virtanen, J., Jääskeläinen, I. P., et al. (1998). Processing of Novel Sounds and Frequency Changes in the Human Auditory Cortex: Magnetoencephalographic Recordings. *Psychophysiology, 35*(2), 211-224.

Allen, J. S., & Miller, J. L. (1999). Effects of Syllable-Initial Voicing and Speaking Rate on the Temporal Characteristics of Monosyllabic Words. *Journal of the Acoustical Society of America, 106*(4), 2031-2039.

Archangeli, D. (1988). Aspects of Underspecification Theory. *Phonology, 5*(2), 183-207.

Assadollahi, R., & Pulvermüller, F. (2003). Early Influences of Word Length and Frequency: A Group Study Using MEG. *NeuroReport, 14*(8), 1183-1187.

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). Mismatch Negativity (MMN): An Objective Measure of Sensory Memory and Long-Lasting Memories During Sleep. *International Journal of Psychophysiology, 46*(3), 215-225.

Aulanko, R., Hari, R., Lounasmaa, O., Näätänen, R., & Sams, M. (1993). Phonetic Invariance in the Human Auditory Cortex. *Neuroreport, 4*(12), 1356-1359.

Bendor, D., & Wang, X. (2005). The Neuronal Representation of Pitch in Primate Auditory Cortex. *Nature, 436*, 1161-1165.

Boersma, P. (2001). Praat, a System for Doing Phonetics by Computer. *Glot International, 5*(9/10), 341-345.

Bonte, M., Valente, G., & Formisano, E. (2009). Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations. *Journal of Neuroscience, 29*(6), 1699-1706.

Böttcher-Gandor, C., & Ullsperger, P. (1992). Mismatch Negativity in Event-Related Potentials to Auditory Stimuli as a Function of Varying Interstimulus Interval. *Psychophysiology, 29*(5), 546-550.

Bradlow, A. R. (1995). A Comparative Study of English and Spanish Vowels. *Journal of the Acoustical Society of America, 97*(3), 1916-1924.

Broad, D. J., & Wakita, H. (1977). Piecewise-Planar Representation of Vowel Formant Frequencies. *Journal of the Acoustical Society of America, 62*(6), 1467-1473.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge, UK: Cambridge University Press.

Cedolin, L., & Delgutte, B. (2005). Pitch of Complex Tones: Rate-Place and Interspike Interval Representations in the Auditory Nerve. *Journal of Neurophysiology, 94*, 347-362.

Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural Response Correlates of Detection of Monaurally and Binaurally Created Pitches in Humans. *Cerebral Cortex, 16*(6), 835-848.

de Cheveigné, A., & Simon, J. Z. (2007). Denoising Based on Time-Shift PCA. *Journal of Neuroscience Methods, 165*, 297-305.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Claes, T., Dologlou, I., ten Bosch, L., & van Compernolle, D. (1998). A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition. *IEEE Transactions on Speech and Audio Processing, 6*(6), 549-557.

Cowan, N., Winkler, I., Teder, W., & Näätänen, R. (1993). Memory Prerequisites of Mismatch Negativity in the Auditory Event-Related Potential (ERP). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 909-921.

Csépe, V. (1995). On the Origin and Development of the Mismatch Negativity. *Ear and Hearing, 16*(1), 91-104.

Csépe, V., Karmos, G., & Molnár, M. (1987). Evoked Potential Correlates of Stimulus Deviance During Wakefulness and Sleep in Cat--Animal Model of

Mismatch Negativity. *Electroencephalography and Clinical Neurophysiology, 66*(6), 571-578.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing Speech Sounds: Top-Down Influences on the Interface between Audition and Speech Perception. *Hearing Research, 229*(1-2), 132-147.

Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological Correlates of Phonological Processing: A Cross-Linguistic Study. *Journal of Cognitive Neuroscience, 12*(4), 635-647.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An Experimental Study of the Acoustic Determinants of Vowel Color: Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns. *Word, 8*(3), 195-210.

Deng, L., & O'Shaughnessy, D. (2003). *Speech Processing: A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker, Inc.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology, 55*, 149-179.

Diesch, E., Eulitz, C., Hampson, S., & Ross, B. (1996). The Neurotopography of Vowels as Mirrored by Evoked Magnetic Field Measurements. *Brain and Language, 53*(2), 143-168.

Disner, S. F. (1980). Evaluation of Vowel Normalization Procedures. *Journal of the Acoustical Society of America, 67*(1), 253-261.

Eimas, P. D. (1985). The Perception of Speech in Early Infancy. *Scientific American, 252*(1), 46-52.

Elberling, C., Bak, C., Kofoed, B., Lebech, J., & Saermark, K. (1981). Auditory Magnetic Fields from the Human Cortex: Influence of Stimulus Intensity. *Scandanavian Audiology, 10*(3), 203-207.

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. *Journal of Neuroscience, 15*(4), 2748-2755.

Eulitz, C., & Lahiri, A. (2004). Neurobiological Evidence for Abstract Phonological Representations in the Mental Lexicon During Speech Recognition. *Journal of Cognitive Neuroscience, 16*(4), 577-583.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fischer, C., Morlet, D., & Giard, M. (2000). Mismatch Negativity and N100 in Comatose Patients. *Audiology & Neuro-otology, 5*(3-4), 192-197.

Fitch, R. H., Miller, S., & Tallal, P. (1997). Neurobiology of Speech Perception. *Annual Review of Neuroscience, 20*, 351-353.

Fitch, W. T., & Giedd, J. (1999). Morphology and Development of the Human Vocal Tract: A Study Using Magnetic Resonance Imaging. *Journal of the Acoustical Society of America, 106*(3), 1511-1522.

Flagg, E. J., Cardy, J. E. O., & Roberts, T. P. L. (2005). MEG Detects Neural Consequences of Anomalous Nasalization in Vowel-Consonant Pairs. *Neuroscience Letters, 397*(3), 263-269.

Formisano, E., de Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science, 322*(5903), 970-973.

Forss, N., Mäkelä, J. P., McEvoy, L., & Hari, R. (1993). Temporal Integration and Oscillatory Responses of the Human Auditory Cortex Revealed by Evoked Magnetic Fields to Click Trains. *Hearing Research, 68*(1), 89-96.

Fowler, C. A. (1986). An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective. *Journal of Phonetics, 14*, 3-28.

Fowler, C. A., & Brown, J. M. (2000). Perceptual Parsing of Acoustic Consequences of Velum Lowering from Information for Vowels. *Perception & Psychophysics, 62*, 21-32.

Fox, R. A., Jacewicz, E., & Feth, L. L. (2008). Spectral Integration of Dynamic Cues in the Perception of Syllable Initial Stops. *Phonetica, 65*(1-2), 19-44.

Friedrich, C. K., Eulitz, C., & Lahiri, A. (2006). Not Every Pseudoword Disrupts Word Recognition. *Behavioral and Brain Functions, 2*, 36.

Frye, R. E., Fisher, J. M., Coty, A., Zarella, M., Liederman, J., & Halgren, E. (2007). Linear Coding of Voice Onset Time. *Journal of Cognitive Neuroscience, 19*(9), 1476-1487.

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. *European Journal of Neuroscience, 18*, 432-440.

Fujisaki, H., & Kawashima, T. (1968). The Role of Pitch and Higher Formants in the Perception of Vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16*(1), 73-77.

Gaeta, H., Friedman, D., Ritter, W., & Cheng, J. (2001). The Effect of Perceptual Grouping on the Mismatch Negativity. *Psychophysiology, 38*(2), 316-324.

Gage, N. M., Poeppel, D., Roberts, T. P. L., & Hickok, G. (1998). Auditory Evoked M100 Reflects Onset Acoustics of Speech Sounds. *Brain Research, 814*, 236-239.

Gage, N. M., & Roberts, T. P. L. (2000). Temporal Integration: Reflections in the M100 of the Auditory Evoked Field. *NeuroReport, 11*(12), 2723-2726.

Gage, N. M., Roberts, T. P. L., & Hickok, G. (2006). Temporal Resolution Properties of Human Auditory Cortex: Reflections in the Neuromagnetic Auditory Evoked M100 Component. *Brain Research, 1069*(1), 166-171.

Ganong, L. H. (1980). Phonetic Categorization in Auditory Word Perception. *Journal of Experimental Psychology: Human Perception and Performace, 6*, 110-125.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes, 12*(5/6), 613-656.

Gaskell, M. G., & Snoeren, N. D. (2008). The Impact of Strong Assimilation on the Perception of Connected Speech. *Journal of Experimental Psychology: Human Perception and Performance, 34*(6), 1632-1647.

Geshwind, N. (1970). The Organization of Language and the Brain: Language Disorders after Brain Damage Help in Elucidating the Neural Basis of Verbal Behavior. *Science, 170*, 940-944.

Giard, M. H., Lavikahen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J., et al. (1995). Separate Representation of Stimulus Frequency, Intensity, and Duration in Auditory Sensory Memory: An Event-Related Potential and Dipole-Model Analysis. *Journal of Cognitive Neuroscience, 7*(2), 133-143.

Goldinger, S. D. (1996a). Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1996b). Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review, 105*(2), 251-279.

Goldstein, J. L. (1973). An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones. *Journal of the Acoustical Society of America, 54*(6), 1496-1516.

Gomes, H., Ritter, W., & Vaughan, H. G. (1995). The Nature of Preattentive Storage in the Auditory System. *Journal of Cognitive Neuroscience, 7*(1), 81-94.

Govindarajan, K. K., Phillips, C., Poeppel, D., Roberts, T. P. L., & Marantz, A. (1998). Latency of MEG M100 Response Indexes First Formant Frequency. *Journal of the Acoustical Society of America, 103*(5), 2982-2983.

Gow, D. W. (2001). Assimilation and Anticipation in Continuous Spoken Word Recognition. *Journal of Memory and Language, 45*, 133-159.

Greenberg, S. (1999). Speaking in Shorthand: A Syllable-Centric Perspective for Understanding Pronunciation Variation. *Speech Communication, 29*(2-4), 159-176.

Greenberg, S. (2006). A Multi-Tier Framework for Understanding Spoken Language. In S. Greenberg & W. A. Ainsworth (Eds.), *Listening to Speech: An Auditory Perspective* (pp. 411-433). Mahwah, NJ: Erlbaum.

Guenther, F. H. (2002). Neural Control of Speech Movements. In A. Meyer & N. Schiller (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Berlin: Mouton de Gruyter.

Haber, R. N. (1969). *Information-Processing Approaches to Visual Perception*. New York: Holt, Rinehart, & Winston.

Hacquard, V., Walter, M. A., & Marantz, A. (2007). The Effects of Inventory on Vowel Perception in French and Spanish: An MEG Study. *Brain and Language, 100*(3), 295-300.

Halberstam, B., & Raphael, L. J. (2004). Vowel Normalization: The Role of Fundamental Frequency and Upper Formants. *Journal of Phonetics, 32*, 423-434.

Halle, M. (1983). On Distinctive Features and Their Articulatory Implementation. *Natural Language and Linguistic Theory, 1*(1), 91-105.

Halle, M. (1995). Feature Geometry and Feature Spreading. *Linguistic Inquiry, 26*(1), 1-46.

Halle, M. (2002). *From Memory to Speech and Back: Papers on Phonetics and Phonology 1954-2002*. Berlin: Mouton de Gruyter.

Halle, M., & Stevens, K. N. (1959). Analysis by Synthesis. In W. Wathen-Dunn & L. E. Woods (Eds.), *Proceedings of the Seminar on Speech Compression and Processing* (Vol. 2, pp. Paper D7): USAF Camb. Res. Ctr.

Halle, M., & Stevens, K. N. (1962). Speech Recognition: A Model and a Program for Research. *IRE Transactions of the PGIT, IT-8*, 155-159.

Halle, M., & Stevens, K. N. (1991). Knowledge of Language and the Sounds of Speech. In J. Sundberg, L. Nord & R. Carlson (Eds.), *Music, Language, Speech, and Brain*. London: MacMillan Press.

Hari, R., Aittoniemi, K., Järvinen, M. L., Katila, T., & Varpula, T. (1980). Auditory Evoked Transient and Sustained Magnetic Fields of the Human Brain. Localization of Neural Generators. *Experimental Brain Research, 40*(2), 237-240.

Hari, R., Hämäläinen, M., Ilmoniemi, R., Kaukoranta, E., Reinikainen, K., Salminen, J., et al. (1984). Responses of the Primary Auditory Cortex to Pitch Changes in a Sequence of Tone Pips: Neuromagnetic Recordings in Man. *Neuroscience Letters, 50*(1-3), 127-132.

Hari, R., Levänen, S., & Raij, T. (2000). Timing of Human Cortical Functions During Cognition. *Trends in Cognitive Sciences, 4*(12), 455-462.

Hawkins, S. (2003). Roles and Representations of Systematic Fine Phonetic Detail in Speech Understanding *Journal of Phonetics, 31*(3-4), 373-405.

Heffner, H., & Whitfield, I. C. (1976). Perception of the Missing Fundamental by Cats. *Journal of the Acoustical Society of America, 59*(4), 915-919.

Hickok, G., & Poeppel, D. (2000). Towards a Functional Neuroanatomy of Speech Perception. *Trends in Cognitive Sciences, 4*(4), 131-138.

Hickok, G., & Poeppel, D. (2007). The Cortical Organization of Speech Processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Hillenbrand, J. M., & Gayvert, R. T. (1993). Identification of Steady-State Vowels Synthesized from the Peterson and Barney Measurements. *Journal of the Acoustical Society of America, 94*(2), 668-674.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic Characteristics of American English Vowels. *Journal of the Acoustical Society of America, 97*(5), 3099-3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of Resynthesized /Hvd/ Utterances: Effects of Formant Contour. *Journal of the Acoustical Society of America, 105*(6), 3509-3523.

Hintzman, D. L. (1986). "Schema Astraction" In a Multiple-Trace Model. *Psychological Review, 93*, 411-428.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of Children, Women, and Men: The Effects of Vocal Intensity Variation. *Journal of the Acoustical Society of America, 106*(3), 1532-1542.

Irino, T., & Patterson, R. D. (2002). Segregating Information About the Size and Shape of the Vocal Tract Using a Time-Domain Auditory Model: The Stabilised Wavelet-Mellin Transform. *Speech Communication, 36*(3/4), 181-203.

Ives, D. T., Smith, D. R. R., & Patterson, R. D. (2005). Discrimination of Speaker Size from Syllable Phrases. *Journal of the Acoustical Society of America, 118*(6), 3816-3822.

Jackson, A., & Morton, J. (1984). Facilitation of Auditory Recognition. *Memory and Cognition, 12*(6), 568-574.

Jacobson, G. P., Lombardi, D. M., Gibbens, N. D., Ahmad, B. K., & Newman, C. W. (1992). The Effects of Stimulus Frequency and Recording Site on the Amplitude and Latency of Multichannel Cortical Auditory Evoked Potential (Caep) Component N1. *Ear and Hearing, 13*(5), 300-306.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic Cognition: Memory, Perception and Concept Formation. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 1-47). New York: Academic Press.

Jakobson, R. C., Fant, G. M., & Halle, M. (1952). *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

Javitt, D. C., Schroeder, C. E., Steinschneider, M., Arezzo, J. C., & Vaughan, H. G., Jr. (1992). Demonstration of Mismatch Negativity in the Monkey. *Electroencephalography and Clinical Neurophysiology, 83*(1), 87-90.

Javitt, D. C., Steinschneider, M., Schroder, C. E., Vaughan, H. G., & Arezzo, J. C. (1994). Detection of Stimulus Deviance within Primate Primary Auditory Cortex: Intracortical Mechanisms of Mismatch Negativity (MMN) Generation. *Brain Research, 667*(2), 192-200.

Johnson, K. (1997). Speech Perception without Speaker Normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145-165). San Diego, CA: Academic Press.

Johnson, K. (2005). Speaker Normalization in Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363-389). Oxford: Blackwell Publishers.

Joos, M. (1948). Acoustic Phonetics. *Language, 24*(2), 5-136.

Jusczyk, P. W., & Luce, P. A. (2002). Speech Perception and Spoken Word Recognition: Past and Present. *Ear and Hearing, 23*(1), 2-40.

Kazanina, N., Phillips, C., & Idsardi, W. J. (2006). The Influence of Meaning on the Perception of Speech Sounds. *Proceedings of the National Academy of Sciences of the United States of America, 103*(3), 11381-11386.

Kekoni, J., Hämäläinen, H., Saarinen, M., Gröhn, J., Reinikainen, K., Lehtokoski, A., et al. (1997). Rate Effect and Mismatch Responses in the Somatosensory System: ERP-Recordings in Humans. *Biological Psychology, 46*(2), 125-142.

Kiefte, M. (2005). Production and Perception of Whispered Vowels. *Journal of the Acoustical Society of America, 118*(3), 1933.

Klatt, D. H. (1989). Review of Selected Models of Speech Perception. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process* (pp. 169-226). Cambridge, MA: MIT Press.

Koenig, W. (1949). *Bell Laboratories Record, 27*, 299-301.

Korzyukov, O. A., Winkler, I., Gumenyuk, V. I., & Alho, K. (2003). Processing Abstract Auditory Features in the Human Auditory Cortex. *Neuroimage, 20*(4).

Krauel, K., Schott, P., Sojka, B., Pause, B. M., & Fersti, R. (1999). Is There a Mismatch Negativity Analogue in the Olfactory Event-Related Potential? *Journal of Psychophysiology, 13*(1), 49-55.

Kuhl, P. K. (1979). Speech Perception in Early Infancy: Perceptual Constancy for Spectrally Dissimilar Vowel Categories. *Journal of the Acoustical Society of America, 66*(6), 1668-1679.

Kuhl, P. K. (1983). Perception of Auditory Equivalence Classes for Speech in Early Infancy. *Infant Behavior & Development, 6*(2-3), 263-285.

Kuhl, P. K. (2004). Early Language Acquisition: Cracking the Speech Code. *Nature Neuroscience, 5*, 831-843.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science, 255*, 606-608.

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *Journal of the Acoustical Society of America, 29*(1), 98-104.

Lahiri, A. (2007). Non-Equivalence between Phonology and Phonetics. *Proceedings of ICPhS XVI*, 31-34.

Lahiri, A., & Marslen-Wilson, W. (1991). The Mental Representation of Lexical Form: A Phonological Approach to the Recognition Lexicon. *Cognition, 38*, 245-294.

Lahiri, A., & Reetz, H. (2002). Underspecified Recognition. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 637-675). Berlin: Mouton de Gruyter.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A Cortical Network for Semantics: (De)Constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920-933.

Liberman, A. M. (1996). *Speech: A Special Code*. Cambridge, MA: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the Speech Code. *Psychological Review, 74*(6), 431-461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some Cues for the Disctinction between Voiced and Voiceless Stops in Initial Position. *Language and Speech, 1*(3), 153-167.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The Discrimination of Relative Onset-Time Patterns of the Components of Certain Speech and Nonspeech Patterns. *Journal of Experimental Psychology, 61*(5), 379-388.

Liberman, A. M., & Mattingly, I. G. (1985). The Motor Theory of Speech Perception Revised. *Cognition, 21*, 1-36.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural Substrates of Phonemic Perception. *Cerebral Cortex, 15*(10), 1621-1631.

Liederman, J., Frye, R., Fisher, J. M., Greenwood, K., & Alexander, R. (2005). A Temporally Dynamic Context Effect That Disrupts Voice Onset Time Discrimination of Rapidly Successive Stimuli. *Psychonomic Bulletin & Review, 12*(2), 380-386.

Lisker, L., & Abramson, A. S. (1964). A Cross-Linguistic Study of Voicing in Initial Stops: Acoustical Measurements. *Word, 20*, 384-422.

Lloyd, R. J. (1890). Speech Sounds: Their Nature and Causation. *Phonetische Studien, 3*, 251-278.

Lombardi, L. (1995). Dahl's Law and Privative Voice. *Linguistic Inquiry, 26*, 356-372.

Lounasmaa, O. V., Hämäläinen, M., Hari, R., & Salmelin, R. (1996). Information Processing in the Human Brain: Magnetoencephalographic Approach. *Proceedings of the National Academy of Sciences, 93*, 8809-8815.

Luce, P. A., & Large, N. R. (2001). Phonotactics, Density, and Entropy in Spoken Word Recognition *Language and Cognitive Processes, 16*(5-6), 565-581.

Lütkenhöner, B., & Steinsträter, O. (1998). High-Precision Neuromagnetic Study of the Functional Organization of the Human Auditory Cortex. *Audiology and Neuro-Otology, 3*(2-3), 191-213.

Maekawa, T., Goto, Y., Kinukawa, N., Taniwaki, T., Kanba, S., & Tobimatsu, S. (2005). Functional Characterization of Mismatch Negativity to a Visual Stimulus. *Clinical Neurophysiology, 116*(10), 2932-2402.

Maiste, A. C., Wiens, A. S., Hunt, M. J., Scherg, M., & Picton, T. W. (1995). Event-Related Potentials and the Categorical Perception of Speech Sounds. *Ear and Hearing, 16*(1), 68-90.

Marslen-Wilson, W., & Warren, P. (1994). Levels of Perceptual Representation and Process in Lexical Access: Words, Phonemes, and Features. *Psychological Review, 101*(4), 653-675.

Matsuwaki, Y., Nakajima, T., Ookushi, T., Iimura, J., Kunou, K., Nakagawa, M., et al. (2004). Evaluation of Missing Fundamental Phenomenon in the Human Auditory Cortex. *Auris Nasus Larynx, 31*, 208-211.

McCarthy, J. J. (1988). Feature Geometry and Dependency: A Review. *Phonetica, 45*, 84-108.

McClelland, J. L., & Elman, J. L. (1986). The Trace Model of Speech Perception. *Cognitive Psychology, 18*, 1-86.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science, 30*(6), 1113-1126.

Meddis, R., & O'Mard, L. P. (2006). Virtual Pitch in a Computational Physiological Model. *Journal of the Acoustical Society of America, 120*(6), 3861-3869.

Mester, R. A., & Itô, J. (1989). Feature Predictability and Underspecification: Palatal Prosody in Japanese Mimetics. *Language, 65*(2), 258-293.

Miller, J. D. (1989). Auditory-Perceptual Interpretation of the Vowel. *Journal of the Acoustical Society of America, 85*(5), 2114-2134.

Mitterer, H., & Blomert, L. (2003). Coping with Phonological Assimilation in Speech Perception: Evidence for Early Composition. *Perception & Psychophysics, 65*(6), 956-969.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An Effect of Linguistic Experience: The Discrimination of [R] and [L] by Native Speakers of Japanese and English. *Perception & Psychophysics, 18*(5), 331-340.

Monahan, P. J., de Souza, K., & Idsardi, W. J. (2008). Neuromagnetic Evidence for Early Auditory Restoration of Fundamental Pitch. *PLoS One, 3*(8), e2900.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some Effects of Variability on Spoken Word Recognition. *Journal of the Acoustical Society of America, 85*(1), 365-378.

Näätänen, R. (1992). *Attention and Brain Function*. Mahwah, NJ: Lawrence Erlbaum Associates.

Näätänen, R. (2001). The Perception of Speech Sounds by the Human Brain as Reflected by Mismatch Negativity (MMN) and Its Magnetic Equivalent (MMNm). *Psychophysiology, 38*(1), 1-21.

Näätänen, R., & Alho, K. (1995). Generators of Electrical and Magnetic Mismatch Responses in Humans. *Brain Topography, 7*(4), 315-320.

Näätänen, R., & Alho, K. (1997). Mismatch Negativity: The Measure for Central Sound Representation Accuracy. *Audiology and Neuro-Otology, 2*(5), 341-353.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-Specific Phoneme Representations Revealed by Electric and Magnetic Brain Responses. *Nature, 385*, 432-434.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The Mismatch Negativity (MMN) in Basic Research of Central Auditory Processing: A Review. *Clinical Neurophysiology, 118*(12), 2544-2590.

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology, 24*(4), 375-425.

Näätänen, R., Sams, M., & Alho, K. (1986). Mismatch Negativity: An ERP Sign of Cerebral Mismatch Process. In R. Zappoli & F. Denoth (Eds.), *Cerebral Psychophysiology: Studies in Event Related Potentials* (pp. 174-180). Amsterdam: Elsevier.

Nashida, T., Yabe, H., Sato, Y., Hiruma, T., Sutoh, T., Shinozaki, N., et al. (2000). Automatic Auditory Information Processing in Sleep. *Sleep, 15*(23), 821-828.

Nearey, T. M. (1989). Static, Dynamic, and Relational Properties in Vowel Perception. *Journal of the Acoustical Society of America, 85*(5), 2088-2113.

Norris, D. (1994). Shortlist: A Continuous Model of Continuous Speech Recognition. *Cognition, 52*, 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging Information in Speech Recognition: Feedback Is Never Necessary. *Behavioral and Brain Sciences, 23*, 299-370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual Learning in Speech. *Cognitive Psychology, 47*, 204-238.

Obleser, J., & Eisner, F. (2009). Pre-Lexical Abstraction of Speech in the Auditory Cortex. *Trends in Cognitive Sciences, 13*(1), 14-19.

Obleser, J., Lahiri, A., & Eulitz, C. (2004). Magnetic Brain Response Mirrors Extraction of Phonological Features from Spoken Vowels. *Journal of Cognitive Neuroscience, 16*(1), 31-39.

Oldfield, R. C. (1971). Assessment and Analysis of Handedness: Edinburgh Inventory. *Neuropsychologia, 9*(1), 97-113.

Pantev, C., Hoke, M., Lehnertz, K., & Lütkenhöner, B. (1989). Neuromagnetic Evidence of an Amplitopic Organization of the Human Auditory Cortex. *Electroencephalography and Clinical Neurophysiology, 72*(3), 225-231.

Pantev, C., Hoke, M., Lütkenhöner, B., & Lehnertz, K. (1989). Tonotopic Organization of the Auditory Cortex: Pitch Versus Frequency Representation. *Science, 246*(4929), 486-488.

Pause, B. M., & Krauel, K. (2000). Chemosensory Event-Related Potentials (CSERP) as a Key to the Psychology of Odors. *International Journal of Psychophysiology, 36*(2), 105-122.

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The Acoustic Bases for Gender Identification from Children's Voices. *Journal of the Acoustical Society of America, 109*(6), 2988-2998.

Peterson, G. E. (1951). The Phonetic Value of Vowels. *Language, 27*, 541-553.

Peterson, G. E. (1961). Parameters of Vowel Quality. *Journal of Speech and Hearing Research, 4*(1), 10-29.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184.

Phillips, C. (2001). Levels of Representation in the Electrophysiology of Speech Perception. *Cognitive Science, 25*, 711-731.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory Cortex Accesses Phonological Categories: An MEG Mismatch Study. *Journal of Cognitive Neuroscience, 12*(6), 1038-1055.

Picton, W., Woods, D. L., Baribeau-Braun, J., & Healey, T. M. (1976). Evoked Potential Audiometry. *Journal of Otolaryngology, 6*(2), 90-119.

Pierrehumbert, J. B. (2002). Word-Specific Phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101-139). Berlin: Mouton de Gruyter.

Pisoni, D. B. (1997). Some Thoughts On "Normalization" In Speech Perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 9-31). San Diego, CA: Academic Press.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-Phonetic Representations in Word Recognition. *Cognition, 25*(1), 21-52.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech Perception at the Interface of Neurobiology and Linguistics: Prospects and Problems. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*, 1071-1086.

Poeppel, D., & Monahan, P. J. (2008). Speech Perception: Cognitive Foundations and Cortical Implementation. *Current Directions in Psychological Science, 17*(2), 80-85.

Poeppel, D., Phillips, C., Yellin, E., Rowley, H. A., Roberts, T. P. L., & Marantz, A. (1997). Processing of Vowels in Supratemporal Auditory Cortex. *Neuroscience Letters, 221*, 145-148.

Potter, R. K., & Steinberg, J. C. (1950). Toward the Specification of Speech. *Journal of the Acoustical Society of America, 22*(6), 807-820.

Pulvermüller, F., & Shtyrov, Y. (2006). Language Outside the Focus of Attention: The Mismatch Negativity as a Tool for Studying Higher Cognitive Processes. *Progress in Neurobiology, 79*(1), 49-71.

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology, 18*(1), 10-30.

Pylyshyn, Z. W. (1985). *Computation and Cognition*. Cambridge, MA: MIT Press.

Ritter, W., Gomes, H., Cowan, N., Sussman, E., & Vaughan, H. G., Jr. (1998). Reactivation of a Dormant Representation of an Auditory Stimulus Feature. *Journal of Cognitive Neuroscience, 10*(5), 605-614.

Ritter, W., Paavilainen, P., Lavikainen, J., Reinikainen, K., Alho, K., Sams, M., et al. (1992). Event-Related Potentials to Repetition and Change of Auditory Stimuli. *Electroencephalography and Clinical Neurophysiology, 83*(5), 306-321.

Ritter, W., & Ruchkin, D. S. (1992). A Review of Event-Related Potential Components Discovered in the Context of Studying P3. In D. Friedman & G. Bruder (Eds.), *Psychophysiology and Experimental Psychopathology: A Tribute to Samuel Sutton* (pp. 1-32). New York: Annals of the New York Academy of Sciences.

Roberts, T. P. L., Ferrari, P., & Poeppel, D. (1998). Latency of Evoked Neuromagnetic M100 Reflects Perceptual and Acoustic Stimulus Attributes. *Neuroreport, 5*(9), 3265-3269.

Roberts, T. P. L., Ferrari, P., Stufflebeam, S. M., & Poeppel, D. (2000). Latency of the Auditory Evoked Neuromagnetic Field Components: Stimulus Dependence and Insights toward Perception. *Journal of Clinical Neurophysiology, 17*(2), 114-129.

Roberts, T. P. L., Flagg, E. J., & Gage, N. M. (2004). Vowel Categorization Induces Departure of M100 Latency from Acoustic Prediction. *Neuroreport, 15*(10), 1679-1682.

Roberts, T. P. L., & Poeppel, D. (1996). Latency of Auditory Evoked M100 as a Function of Tone Frequency. *NeuroReport, 7*(6), 1138-1140.

Rosner, B. S., & Pickering, J. B. (1994). *Vowel Perception and Production*. Oxford: Oxford University Press.

Sams, M., Hari, R., Rif, J., & Knuutila, J. (1993). The Human Auditory Sensory Memory Trace Persists About 10 Sec: Neuromagnetic Evidence. *Journal of Cognitive Neuroscience, 5*(3), 363-370.

Sams, M., Paavilainen, P., Alho, K., & Näätänen, R. (1985). Auditory Frequency Discrimination and Event-Related Potentials. *Electroencephalography and Clinical Neurophysiology, 62*(6), 437-448.

Scherg, M., Vajsar, J., & Picton, T. W. (1989). A Source Analysis of the Late Human Auditory Evoked Potentials. *Journal of Cognitive Neuroscience, 1*(4), 336-355.

Schouten, J. F. (1970). The Residue Revisited. In R. Plomp & G. F. Smoorenberg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing* (pp. 41-54). Leiden, The Netherlands: Sijthoff.

Schouten, J. F., Ritsma, H. J., & Cordozo, B. L. (1962). Pitch of the Residue. *Journal of the Acoustical Society of America, 34*(9B), 1418-1424.

Scott, S. K., & Johnsrude, I. S. (2003). The Neuroanatomical and Functional Organization of Speech Perception. *Trends in Neurosciences, 26*(2), 100-107.

Sculthorpe, L. D., Ouellet, D. R., & Campbell, K. B. (2009). MMN Elicitation During Natural Sleep to Violations of an Auditory Pattern. *Brain Research, Epub Ahead of Print*, doi:10.1016/j.brainres.2009.1006.1013

Shamma, S., & Klein, D. (2000). The Case of the Missing Pitch Templates: How Harmonic Templates Emerge in the Early Auditory System. *Journal of the Acoustical Society of America, 107*(5), 2631-2644.

Sharma, A., & Dorman, M. F. (1999). Cortical Auditory Evoked Potential Correlates of Categorical Perception of Voice-Onset Time. *Journal of the Acoustical Society of America, 106*(2), 1078-1083.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic Correlates of Cross-Language Phonetic Perception. *Journal of the Acoustical Society of America, 107*(5), 2697-2703.

Sharma, A., Kraus, N., McGee, T., Carrell, T., & Nicol, T. (1993). Acoustic Versus Phonetic Representation of Speech as Reflected by the Mismatch Negativity Event-Related Potential. *Electroencephalography and Clinical Neurophysiology, 88*(1), 64-71.

Shinozaki, N., Yabe, H., Sutoh, T., Hiruma, T., & Kaneko, S. (1998). Somatosensory Automatic Responses to Deviant Stimuli. *Cognitive Brain Research, 7*(2), 165-171.

Shtyrov, Y., & Pulvermüller, F. (2002). Neurophysiological Evidence of Memory Traces for Words in the Human Brain. *Neuroreport, 25*(13), 521-525.

Shtyrov, Y., Pulvermüller, F., Näätänen, R., & Ilmoniemi, R. J. (2003). Grammar Processing Outside the Focus of Attention: An MEG Study. *Journal of Cognitive Neuroscience, 15*(8), 1195-1206.

Silverman, D. (2006). *A Critical Introduction to Phonology: Of Sound, Mind, and Body*. London: Continuum Books.

Simpson, T. P., Manara, A. R., Kane, N. M., Burton, R. L., Rowlands, C. A., & Butler, S. R. (2002). Effect of Propofol Anaesthesia on the Event-Related Potential Mismatch Negativity and the Auditory-Evoked Potential N1. *British Journal of Anaesthesia, 89*(3), 382-388.

Sinex, D. G., & McDonald, L. P. (1988). Average Discharge Rate Representation of Voice Onset Time in the Chinchilla Auditory Nerve. *Journal of the Acoustical Society of America, 83*(5), 1817-1827.

Slawson, A. W. (1968). Vowel Quality and Musical Timbre as Functions of Spectrum Envelopes and Fundamental Frequency. *Journal of the Acoustical Society of America, 43*(1), 87-101.

Smith, D. R. R., & Patterson, R. D. (2005). The Interaction of Glottal-Pulse Rate and Vocal-Tract Length in Judgements of Speaker Size, Sex, and Age. *Journal of the Acoustical Society of America, 118*(5), 3177-3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irnio, T. (2005). The Processing and Perception of Size Information in Speech Sounds. *Journal of the Acoustical Society of America, 117*(1), 305-318.

Smoorenburg, G. F. (1970). Pitch Perception of Two-Frequency Stimuli. *Journal of the Acoustical Society of America, 48*(4B), 924-942.

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan, H. G., Jr. (1995). Physiologic Correlates of the Voice Onset Time Boundary in Primary Auditory Cortex (A1) of the Awake Monkey: Temporal Response Patterns. *Brain and Language, 48*(3), 326-340.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2002). Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features. *Journal of the Acoustical Society of America, 111*(4), 1872-1891.

Stevens, K. N., & Bickley, C. (1991). Constraints among Parameters Simplify Control of Klatt Formant Synthesizer. *Journal of Phonetics, 19*, 161-174.

Stevens, K. N., & Halle, M. (1967). Remarks on Analysis by Synthesis and Distinctive Features. In W. Wathen-Dunn (Ed.), *Models for the Perception of Speech and Visual Form* (pp. 88-102). Cambridge, MA: MIT Press.

Stevens, S. S., & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology, 53*(3), 329-353.

Strange, W. (1989). Evolving Theories of Vowel Perception. *Journal of the Acoustical Society of America, 85*(5), 2081-2087.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic Specification of Coarticulated Vowels. *Journal of the Acoustical Society of America, 74*(3), 695-705.

Studdert-Kennedy, M. (1976). Speech Perception. In N. J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 243-293). New York: Academic Press.

Studdert-Kennedy, M. (1980). Speech Perception. *Language and Speech, 23*(1), 45-66.

Stufflebeam, S. M., Poeppel, D., Rowley, H. A., & Roberts, T. P. L. (1998). Peri-Threshold Encoding of Stimulus Frequency and Intensity in the M100 Latency. *Neuroreport, 5*(9), 91-94.

Sussman, H. M. (2000). Phonemic Representation: A Twenty-First Century Challenge. *Brain and Language, 71*(1), 237-240.

Syrdal, A. K., & Gopal, H. S. (1986). A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels. *Journal of the Acoustical Society of America, 79*(4), 1086-1100.

Tales, A., Newton, P., Troscianko, T., & Butler, S. (1999). Mismatch Negativity in the Visual Modality. *Neuroreport, 8*(10), 3363-3367.

Tavabi, K., Elling, L., Dobel, C., Pantev, C., & Zwitserlood, P. (2009). Effects of Place of Articulation Changes on Auditory Neural Activity: A Magnetoencephalography Approach. *PLoS ONE, 4*(2), e4452.

Terhardt, E. (1974). Pitch, Constancy, and Harmony. *Journal of the Acoustical Society of America, 55*(5), 1061-1069.

Tervaniemi, M., Lehtokoski, A., Sinkkonen, J., Virtanen, J., Ilmoniemi, R. J., & Näätänen, R. (1999). Test-Retest Reliability of Mismatch Negativity for Duration, Frequency and Intensity Changes. *Clinical Neurophysiology, 110*(8), 1388-1393.

Tiitinen, H., Alho, K., Huotilainen, M., Ilmoniemi, R. J., Simola, J., & Näätänen, R. (1993). Tonotopic Auditory Cortex and the Magnetoencephalographic (MEG) Equivalent of the Mismatch Negativity. *Psychophysiology, 30*(5), 537-540.

Tiitinen, H., Mäkelä, A. M., Mäkinen, V., May, P. J., & Alku, P. (2005). Disentangling the Effects of Phonation and Articulation: Hemispheric Asymmetries in the Auditory N1m Response of the Human Brain. *BMC Neuroscience, 6*, 62.

Tiitinen, H., May, P., Reinikainen, K., & Näätänen, R. (1994). Attentive Novelty Detection in Humans Is Governed by Pre-Attentive Sensory Memory. *Nature, 372*(6501), 90-92.

Tomlinson, R. W. W., & Schwartz, D. W. F. (1988). Perception of the Missing Fundamental in Nonhuman Primates. *Journal of the Acoustical Society of America, 84*(2), 560-565.

Tulving, E., & Schacter, D. L. (1990). Priming and Human Memory Systems. *Science, 247*, 301-306.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433-460.

Vanhaudenhuyse, A., Laureys, S., & Perrin, F. (2008). Cognitive Event-Related Potentials in Comatose and Post-Comatose States. *Neurocritical Care, 8*(2), 262-270.

Vasama, J. P., Mäkelä, J. P., Tissari, S. O., & Hämäläinen, M. S. (1995). Effects of Intensity Variation on Human Auditory Evoked Magnetic Fields. *Acta Otolaryngology, 115*(5), 616-621.

Virtanen, J., Ahveninen, J., Ilmoniemi, R. J., Näätänen, R., & Pekkonen, E. (1998). Replicability of MEG and Eeg Measures of the Auditory N1/N1m-Response. *Electroencephalography and Clinical Neurophysiology, 108*(3), 291-298.

Werker, J. F., & Tees, R. C. (1984). Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life. *Infant Behavior & Development, 7*, 49-63.

Winkler, I., Lehtoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., et al. (1999). Pre-Attentive Detection of Vowel Contrasts Utilizes Both Phonetic and Auditory Memory Representations. *Cognitive Brain Research, 7*(3), 357-369.

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., & Näätänen, R. (1990). The Effect of Small Variation of the Frequent Auditory Stimulus on the Event-Related Brain Potential to the Infrequent Stimulus. *Psychophysiology, 27*(2), 228-235.

Winkler, I., Paavilainen, P., & Näätänen, R. (1992). Can Echoic Memory Store Two Traces Simultaneously? A Study of Event-Related Brain Potentials. *Psychophysiology, 29*(3), 337-349.

Winkler, I., Tervaniemi, M., & Näätänen, R. (1997). Two Separate Codes for Missing-Fundamental Pitch in the Human Auditory Cortex. *Journal of the Acoustical Society of America, 102*(2), 1072-1082.

Wise, R., Chollet, F., Hadar, U., Friston, K., Hoffner, E., & Frackowiak, R. (1991). Distribution of Cortical Neural Networks Involved in Word Comprehension and Word Retrieval. *Brain, 114*(4), 1803-1817.

Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal Window of Integration Revealved by MMN to Sound Omission. *Neuroreport, 8*(8), 1971-1974.

Yeung, H. H., & Phillips, C. (2004). *Phonological Features Distinct from Phonemes in Auditory Cortex*. Paper presented at the 11th Annual Meeting of the Cognitive Neuroscience Society.

Zahorian, S. A., & Jagharghi, J. (1993). Spectral-Shape Features Versus Formants as Acoustic Correlates for Vowels. *Journal of the Acoustical Society of America, 94*(4), 1966-1982.

Zwicker, E. (1961). Subdivision of the Audible Frequency Rage into Critical Bands. *The Journal of the Acoustical Society of America, 33*(2), 248.