NEURAL RESPONSES DEMONSTRATE THE DYNAMICITY OF SPEECH PERCEPTION

NEURAL RESPONSES DEMONSTRATE THE DYNAMICITY OF SPEECH PERCEPTION

By SAMANTHA KRAMER, B.A. (Hons.)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2014) Hamilton, Ontario (Cognitive Science of Language)

TITLE: Neural responses demonstrate the dynamicity of speech perception

AUTHOR: Samantha Kramer, B.A. (Hons.) (McMaster University)

SUPERVISORS: Dr. John F. Connolly and Dr. Anna L. Moro

NUMBER OF PAGES: ix, 75

Abstract

Spoken language is produced with a great deal of variability with which listeners must be able to cope. One source of variation is coarticulation, which is due to articulatory planning and transitions between segments. Recently, the temporal features of coarticulation were investigated during a picture/spoken-word matching task by using spliced stimuli carrying either congruent or incongruent subphonemic cues at the CV juncture (Archibald & Joanisse, 2011). ERPs were recorded with attention paid to the *phonological mapping negativity* (PMN) (Connolly & Phillips, 1994; Newman & Connolly, 2004) – a prelexical response sensitive to violations of phonological expectations. Results found that the PMN varied in response to coarticulation violations and concluded that phonetic features in spoken words influence prelexical processing during word recognition. Using a written-/spoken-word paradigm, Arbour, 2012 controlled phonological shape by using onsets that were either fricatives or stops, hypothesizing that coarticulatory information would be differentially processed due to their temporal differences. Findings supported the PMN's sensitivity to coarticulation but also showed that temporal and physical differences between onsets modulated the effect. These results raise the question of whether acoustic distance between vowels will modulate prelexical processing of speech as reflected by the PMN amplitude: the focus of the current study. Words were organized into minimal sets such that all onset/coda combinations appeared with each vowel provided that English words resulted. Vowels were one of /i, u, æ, ɑ/, maximizing acoustic distance (height and backness). Data from 20 subjects indicate that the PMN is sensitive to the *degree of difference* between the

original and post-splice vowels. When the number of distinctive features changing is greater, the result is an earlier, more robust PMN. This suggests that the rate of speech recognition is not static but dynamic, and is dependent on likeness of subphonemic features.

Acknowledgements

I would like to thank my supervisors Dr. John F. Connolly and Dr. Anna L. Moro for their guidance during the completion of this thesis, despite numerous other urgent responsibilities. I would also like to thank Dr. Victor Kuperman for his help with stimuli frequency data and Dr. Elisabet Service for her statistical insights and help with my writing throughout my time in the M.Sc. program.

Thank you to all my fellow graduate students for listening to my ideas, presentations, and frustrations and offering encouragement when it was very much needed. Additional thanks are due to the members of the Language, Memory and Brain Lab for their hours of help with data collection and running participants! Extra-special thanks goes to Richard Mah for being an indispensible source of knowledge, technical support, and obscure internet references.

Finally, I'd like to thank my family and my fiancé for their endless support.

Table of Contents

List of Figures and Tables

1.0 Introduction

The speech stream contains a great deal of variability with which listeners must cope in order to achieve successful communication. One source of variability, coarticulation, is the focus of the present study. Coarticulation refers to the spatial and temporal modification of the articulation of a speech sound. This variation is due to the articulatory and gestural planning of surrounding segments in the speech stream such that featural overlap arises. Transitions are modulated primarily by the features of surrounding phonemes (speech sounds), as well as by speech rate. These modulations are continuous and subtle (Gow & McMurray, 2007).

How listeners process these subtle acoustic changes in the speech stream is a major topic of investigation in the speech perception literature. A number of behavioural studies have investigated the role of coarticulation in speech perception. Two schools of thought have been established regarding the role of subphonemic cues in speech perception. First, some researchers state that all variation in speech is filtered out and that listeners only deal in canonical phonemes, disregarding the inherent variation (Anderson, 1973; Stevens, 2002). Alternatively, a number of empirical studies have supported the idea that coarticulation is used during speech perception and integrated into the processing stream to facilitate word recognition (McQueen, Norris & Cutler, 1999; Gow & McMurray, 2007; Archibald & Joanisse, 2011; Arbour, 2012). These within-category variations, or subphonemic cues, have been shown to be a valuable source of information to listeners, and are the focus of the current study.

1.1 Coarticulation in Production

Coarticulation is an inherent part of speech production. As the vocal tract configuration moves between targets, there is an overlap in articulatory gestures. Coarticulation is bidirectional. Carryover coarticulation moves from left to right and occurs when preceding sounds affect the articulation of subsequent sounds. Anticipatory coarticulation moves from right to left. This type of coarticulation occurs when the articulatory planning of upcoming segments affects the articulation of previous segments (Recasens, Pallarès & Fontdevila, 1997). Thus, the exact vocal tract configuration for any given phoneme is not identical each time it is realized in speech. The exact realization of a phoneme is largely dependent upon the surrounding segments.

This phenomenon was illustrated in Arbour (2012) with the following example. The phoneme /k/ is produced differently in the two words "keep" and "coop" due to anticipatory coarticulation of the following vowel. The /k/ in "keep" is realized as [k^{j}]. This velar stop is palatalized due to being produced adjacent to a high front vowel /i/. The tongue's position when producing a canonical velar [k] carries the feature [+back], but it loses this feature due to anticipating the following vowel, which carries the feature [-back]. The /k/ in "coop" is realized as [k^{w}]. The initial consonant is now produced with lip rounding due to anticipating the upcoming rounded vowel /u/. This differs from the canonical [k] production by acquiring the feature [+round], which is lacking in the canonical configuration (Reetz & Jongman, 2011).

Coarticulation being an inherent property of spoken words warrants its study as a perceptual phenomenon. The role of coarticulation in perception has been studied via

behavioural and neurolinguistic methods. Results of such studies will be outlined in the following section.

1.2 Coarticulation in Perception

Subphonemic cues have been reviewed at length in the phonetics literature. The past literature has seen some debate as to whether listeners discard variance in speech and extract only those consistent features (Stevens, 2002), or whether listeners use perceptual cue variance to their advantage when processing speech (Gow & McMurray, 2007). Due to the phonotactic constraints on languages, there is a limited – albeit large – number of sound sequences legally permissible in any given language. Thus, coarticulatory variation is "lawful" (McQueen & Cutler, 1997). Assuming that listeners encode this lawful variation in their phonemic representations, like they do with allophonic variation, means that these cues could be used to listeners' advantage, facilitating speech processing (Fleming, 1997).

Wright (2004) gives an overview of perceptual cues and their robustness. Wright introduces the idea of robustness in encoding, which is important for speech since it is largely produced in noisy environments. Robust encoding explains the idea that each phoneme's identity must be encoded redundantly, in case some cues are lost to noise. Wright talks about the relationship between sonority and robustness of encoding, saying that a string of obstruents (obstruents are not sonorous) results in weak encoding. Alternating CVCV sequences (vowels are the most sonorous speech sounds) results in strong encoding, making this sequence less likely to be susceptible to noise.

Past behavioural studies in the literature have demonstrated the perceptual effects of coarticulation on listeners. McQueen et al. (1999) conducted a forced-choice phoneme decision task with Dutch speakers in a series of six experiments. In their first experiment, the researchers recorded several Dutch words and nonwords that all ended in a legal vowel-consonant sequence. The final consonant was always a voiceless stop. This gave the researchers experimental tokens such as *sloop* 'pillowcase' and *sloot*, which is a nonword in Dutch. The researchers cross-spliced the final consonant onto other words with otherwise identical phonemic structure, like the two words given above. Tokens were also spliced together with productions of the same word to create identity-spliced tokens, eliminating a splicing confound. This method of cross-splicing created misleading coarticulation in the vowel segment. In the original production of *sloop*, the vowel /o/ carries cues leading to the following /p/. When a /t/ is spliced on, this creates a misleading coarticulatory environment. Participants were told that they would hear an isolated word or nonword over their headphones. They were then instructed to indicate, as quickly as possible, the identity of the final consonant of each word they heard. Participants were given a two-option forced-choice response interface on a computer. The choices for each trial were the pre- and post-splice consonant. Keeping with the following example, the options given to participants would be /p/ and /t/. Reaction times for participants' responses were analyzed to reveal whether there was a measureable effect of coarticulation on response times. If coarticulation cues are ignored by listeners, the expected result would be that there would be no measurable difference between reaction times to identity- and cross-spliced tokens.

Results of the first experiment showed a main effect of lexical status (McQueen et al., 1999). Participants responded more quickly to real words than to nonwords. Further analysis revealed a significant difference in response times between identity- and cross-spliced tokens. Listeners demonstrated significantly longer response latencies when they were presented with cross-spliced tokens. These results support the idea that coarticulation in meaningfully processed by listeners. McQueen et al. (1999) replicated these results by giving participants more phonemes as response options still finding significant differences between identity- and cross-spliced tokens.

To reinforce the results described above, McQueen et al. (1999) ran an additional experiment to test only the vowels used in their previous experiments. Participants heard an isolated vowel sound and were asked to indicate which voiceless consonant originally had followed that vowel using the same forced-choice paradigm employed previously. The goal of this experiment was to show the role of anticipatory coarticulation in the vowel segments. Results revealed an overall response accuracy of 80%. This result shows that there exist enough coarticulatory cues regarding the place of articulation of the upcoming stop to mislead listeners in the case of cross-spliced tokens. These results reinforce the suggestion that coarticulation is not discarded as random, useless variation during speech processing but rather is meaningfully processed by listeners.

Gow (2001; 2002; 2003; Gow & McMurray, 2007) has done extensive research regarding the effect that subphonemic cues have on speech perception with particular reference to the assimilation of coronals (consonants articulated with the apex of the tongue) in connected speech. Assimilation is a process that is comparable to

coarticulation in that a speech sound carries features of its surrounding environment. Conclusions drawn from assimilation can be useful to the study of coarticulation, because both processes involve phonemes taking on features of surrounding sounds.

Gow (2001) introduces the assimilation process using the exemplary phrase "teen player," which is represented underlyingly as /tin plejər/, and in natural speech is realized as [tim plejər]. This change arises through the process of assimilation of the coronal to the following bilabial consonant. It is unclear how listeners identify the first word is 'teen' with a labialized coronal, rather than 'team' with a labial consonant underlying. Gow presents three possible accounts for this perceptive process: tolerance to mismatch, underspecification, and regressive inference.

A tolerance-to-mismatch account of speech perception is one that proposes a tolerance of modifications that arise from speech rate and dialectal variations, among other factors (Marslen-Wilson, 1978; 1993). Support for this account comes from experiments showing that listeners perceive sounds as identical, even if they mismatch in one to two features, when they are found at the end of a word (Cole, 1973; Cole & Jakimik, 1978). This account results in a high number of lexical candidates, due to its lenient feature-matching criteria, but later uses context to disambiguate meaning.

Underspecification accounts of speech perception can be explained by using the example of 'green beans', which is often realized in natural speech as [grim binz] . Gow suggests that the final consonant in 'green' is underspecified for its *place* feature. That is to say that a listener encodes a nasal but does not specify place (i.e. whether it is labial, coronal or velar, since neither *[grim] nor *[griŋ] are lexical competitors). When the

nasal's place of articulation can be explained through the place of the following consonant (i.e. it is labialized due to the labial stop in 'beans'), then listeners tolerate non-exact matching criteria for that lexical item. When feature variations cannot be explained through surrounding context, then stricter feature matching criteria are applied to that item (Lahiri & Marslen-Wilson, 1991).

The third account Gow presents is a regressive inference account that hypothesizes listeners rely wholly on the following context to determine whether phonemes should be processed as they were realized, or whether they have undergone assimilatory processes. By this account, [grim binz] would be processed as 'green' because labialization can be attributed to the labial stop. If a listener heard [grim kajt] (Gow, 2001), they would process the labial nasal as it was produced, since a [k] would not cause labialization.

Gow conducted a series of phoneme monitoring experiments, hypothesizing that reaction times would be quicker if tokens were primed by logical assimilatory processes rather than implausible assimilatory processes, which would increase reaction times. The results of these experiments led Gow to conclude that listeners gain a processing advantage when targets were presented with plausible assimilatory processes (2001, 2002, 2003). This result was reflected in the faster reaction times in the presence of congruent assimilatory processes as compared with slower reaction times when assimilation did not predict the following target. Thus, these results support the idea that coarticulation between segments is meaningfully processed, and can facilitate speech processing.

These results fit within Gow's regressive inference account of speech perception, and was the basis for the subsequent discussion of feature alignment. Gow discusses the proposal that, in natural speech, phonetic features may be associated with more than one segment, resulting in coarticulated or assimilated segments. Aligning features with their intended segments is not a trivial task. Lahiri & Marslen-Wilson (1991) suggest features may be associated to any segment within certain window, namely one that is three segments in length. In this view, assimilation is accomplished through determining the correct feature mapping sequence. Assimilation (and coarticulation) distorts features' canonical mapping, causing listeners to reassign feature mappings until any ambiguity is resolved.

Gow (2003) goes on to suggest, like McQueen et al. (1999) that speakers have an implicit knowledge within their phonological systems of non-linear feature mappings that are the result of phonotactically legal sequences of phonemes. When cues present in a segment are congruent with the upcoming context, this facilitates the processing of the following segments, and allows for disambiguation of ambiguous feature mappings (Smits, 2001). A regressive inference account that includes feature parsing as a component is a bidirectional model of how listeners compensate for assimilation and coarticulation in natural speech.

1.3 Electrophysiological Responses to Coarticulation

The nature of coarticulation is highly variable and temporally impermanent. It is constantly varying , with each phoneme being influenced by previous and subsequent sounds (Farnetani & Recasens, 1997). In this way, it is not a phenomenon that lends

itself well to behavioural studies. Additionally, off-line, behavioural measures only reveal the end result of a potentially multi-step process. Thus, coarticulation can be better studied by an on-line technique with high temporal accuracy that can show processing as it unfolds in real time. Consequently, the effect of subphonemic cues has been studied using neuroimaging techniques, especially electroencephalography.

Electroencephalography (EEG) is a continuous measure of neural activity measured from the scalp. EEG activity and its derivatives such as evoked potentials (EPs) and event-related potentials (ERPs) have been demonstrated to be sensitive to an array of sensory, perceptual and cognitive processes (Kutas & Federmeier, 2011). ERPs are time-based waveforms derived from ongoing EEG that are emitted by an individual and typically reflect a cognitive or cognitively mediated neural activity related to a stimulus event. ERP waveforms linked directly to specific functions (e.g., semantic processing, recognition memory) are referred to as components; thus the N400 component or the P300. Components are identified by their polarity – whether they are a negative-going or positive-going deflection in the waveform – along with the peak latency, which is measured in milliseconds from stimulus onset to the maximum deflection of the response. There are three particular ERP components relevant to the current topic.

The N100 is a negative-going component that peaks approximately 100 ms after stimulus onset, is characterized by a fronto-central distribution, and is generally modality nonspecific reflecting early sensory processing, such as the brightness of a visual stimulus, or the intensity (dB) of an auditory stimulus (Steinhauer & Connolly, 2008).

Following the N100, a positive-going deflection peaking around 200 ms is elicited. This P200 shows a typically central distribution and is often found to be sensitive to stimulus intensity as well (Martin, Tremblay & Stapells, 2007). Together, this early sequence of the N100 and P200 typically reflects auditory stimulus detection in the auditory cortex (Steinhauer & Connolly, 2008; Martin et al., 2007). However, the P200 has been found to be sensitive to intensity within language contexts exemplified by larger amplitudes to stop-burst than to fricative consonants (Arbour, 2012).

The Phonological Mapping Negativity (PMN) is a negative-going waveform that peaks between 230-350 ms post-stimulus, and is found to be fronto-centrally distributed across the scalp (Connolly & Phillips, 1994; Newman & Connolly, 2009). The amplitude of the PMN reflects a violation by the incoming speech signal of a predetermined phonological expectation, set up by some previous prime context ranging from pictures (Connolly et al., 1995) and sentences (Connolly & Phillips, 1994) to phoneme deletion tasks (Newman & Connolly, 2009). Thus, the PMN is elicited when what is perceived does not match what is expected in the speech stream. Connolly & Phillips (1994) showed a double dissociation between the PMN and the N400 (which is most robust in the presence of a semantic anomaly, Kutas & Hillyard, 1980), while other work has demonstrated that the PMN occurs to both words and nonwords (Connolly, Service, D'Arcy, Kujala & Kimmo, 2001; Newman, Connolly, Service & McIvor, 2003). Therefore, the PMN has been described as a distinct reflection of prelexical speech processing. Additionally, Newman et al. (2003) found no significant difference in the PMN's amplitude relating to degree of difference between phonological expectations and

the actual violation, so the PMN was described as an all-or-nothing response rather than a graded response.

In the past, studies such as those above have tested the sensitivity of the PMN to violations between phonemes. Thus, the PMN is well documented as reflecting a between-category violation (Connolly & Phillips, 1994; Connolly et al., 2001; Newman et al., 2003). Archibald & Joanisse (2011) studied the effects of coarticulatory miscues on listeners and whether such coarticulatory mismatches would elicit the PMN in a manner similar to the way that a true phonemic, between-category violation does. They examined three components: the N100, the PMN, and the N400. Archibald & Joanisse (2011) created coarticulatory miscues using a similar splicing method to McQueen et al. (1999) with natural speech stimuli. The splicing process targeted the initial consonant in words with CVC structure. Splicing was done by, for example, replacing the /h/ sound from the word *hoot* with an /h/ sound that came from the word *heat*. Analogous with the previous *coop/keep* example, these two glottal fricatives realized as [h^{w}] and [h^{j}], representing the same underlying phoneme /h/, each set up different expectations of what vowel was to follow due to the process referred to as *anticipatory coarticulation* (Archibald & Joanisse, 2011).

Archibald & Joanisse's (2011) stimuli were 30 imageable words of English that were CVC structured. The initial consonant was the spliced target. Onset consonants were one of the following sounds /f, s, ʃ, ʧ, ʤ, m, n, h/, followed by a vowel and a consonant, creating a monosyllabic closed syllable. Word pairs sharing the same onsets were chosen as stimuli. Onsets were then cross-spliced to create misleading

coarticulatory environments at the CV-juncture. For example, the onset /f/ in *feed* and *food* would be cross-spliced onto the opposite VC sequence, creating /fʲud/ and /fʷid/. Participants were then presented with spoken words containing either congruent or incongruent coarticulatory information while their EEG was recorded.

Archibald & Joanisse (2011) used a picture/spoken-word matching paradigm with the picture creating a context that supported an expectation of what word they would hear subsequently. The spoken word was either a match or a mismatch to the picture on three critical levels: (1) lexical mis/match, (2) phonemic mis/match, and (3) coarticulatory mis/match. The researchers hypothesized that if coarticulation is disregarded by listeners as random noise, then incongruent coarticulatory cues would modulate the N100 response exclusively. Conversely, if coarticulation is being meaningfully processed by listeners then incongruent coarticulatory cues would modulate the PMN response.

Archibald & Joanisse (2011) found no significant difference in N100 amplitude between congruent and incongruent conditions. This lack of modulation suggests that coarticulation is not processed simply as noise in the speech stream. Moreover, they observed similar PMN elicitation to both between-category (phonological) violations and within-category (coarticulatory) violations. These differences in participants' brain potentials show that listeners are processing coarticulation online. Listeners are using these cues to their perceptual advantage, and are led astray in the presence of subphonemic miscues. Archibald & Joanisse (2011) concluded that prelexical processing is more complex than broad phoneme categorization, and that subphonemic cues must be

considered when modeling the early stages of speech perception. These results suggest that the PMN is sensitive to lower-level information than what was reported by Connolly.

Archibald & Joanisse (2011) were the first to demonstrate a PMN to within-category violations, thus necessitating replications and further refinement. As previously mentioned, the stimuli used in their study employed a range of onsets (/f, s, ʃ, ʧ, ʤ, m, n, h/) that are so acoustically and physically different from each other that they cannot be comparable. This wide range of onsets was attributable to the necessity of using only imageable stimuli because of the picture element of the picture/spoken-word matching paradigm. Additionally, the authors claimed that all the stimuli took the phonological shape CVC. This claim is questionable because the list includes words like *news*, which in the local dialect is pronounced /njuz/. Such words take the phonological shape CGVC – with a glide obstructing the consonant-vowel juncture being studied.

In an extension of the Archibald and Joanisse study, Arbour (2012) used onsets consisting of voiceless stops /p, t, k/ and voiceless fricatives /f, s, ʃ, h/ only. Additionally, only the four corner vowels /i, u, æ, ɑ/ were placed in nucleus position, in order to create maximal disparity between tokens. In order to make this refined phonological shape possible, a written-word/spoken-word paradigm was used thus eliminating the imageability problem. The splicing method and the use of congruent and incongruent coarticulatory (mis)cues replicated Archibald & Joanisse (2011). Participants saw a written word on a computer screen that created the expectation or context that was followed by a spoken word that had an appropriate or inappropriate coarticulation cue.

A maximal PMN was elicited in conditions that contained subphonemic miscues providing clear evidence that the use of coarticulatory information in speech recognition varies in strength and timing as a function of onset type (fricative vs. stop) and vowel height (high vs. low). Coarticulatory cues were more readily perceived in spoken words that began with fricatives than stops and subphonemic variations were detected more easily in low vowels than high vowels.

1.4 Models of Spoken Word Recognition

Some fundamentally differing models to describe spoken word recognition have been proposed in the literature. There are many studies that support each model, hence there remains uncertainty about the cognitive processes underlying spoken word recognition. Fundamentally, it is uncertain as to whether top-down information, such as context, influences spoken word recognition, or whether the process is unidirectional, taking only bottom-up information into account.

Top-down models propose that context does indeed play a part in our spoken word comprehension by only considering forms that fit within the established discourse context. Arguably the best know top-down model of speech recognition is the TRACE model (McClelland & Elman, 1986). TRACE proposes that speech sounds are processed through a network of "units" of different classes: acoustic-phonetic, phonological, and lexical. According to this model, these proposed units each have their own activation level and threshold. These activation levels are continuously fluctuating as speech unfolds over time. These fluctuations reflect the listener's hypothesis as to which word they are hearing at the given point in time. The TRACE model allows feedback within

the system, so activation levels of these units are influenced from both directions – bottom-up as well as top-down. For example, lexical expectations (top-down) and the acoustic-phonetic cues (bottom-up) from a speech stream will influence the listener's perception of a phoneme. As each unit reaches its activation threshold, it will stimulate relevant units, and inhibit non-eligible candidate units.

Coarticulation fits into the TRACE model at the level of the proposed acoustic-phonetic, or featural units. TRACE would predict that a mismatch in coarticulatory cues should not fundamentally impede speech recognition, since activation from the lexical level should be sufficiently compensatory. In fact, listeners may not perceive coarticulation when it is there, nor miss it when it is lacking. Results supporting this prediction can be found in a behavioural experiment by Elman & McClelland (1988), which found that listeners perceptually compensate for coarticulation even in its absence. The focus of this study was to investigate whether contextual information aids listeners in their perception of ambiguous speech sounds. They attached an artificial, ambiguous sound (a synthesized fricative between the English /s/ and /ʃ/) to the end of words. The examples given were one of either *Christma_* (taken from a natural speech token of *Christmas*) or *fooli_* (taken from a natural speech token of *foolish*). These words were then followed by an ambiguous sound (a synthetic stop sound between the English /t/ and /k/) at the beginning of *_apes*. The *_apes* tokens were counterbalanced and taken from two natural speech productions of *tapes* and *capes*. Participants were presented with two-word phrases (*Christma_ apes, fooli_ _apes*) that included the two synthesized ambiguous sounds in the spaces left blank, above. After hearing the phrases, participants

were asked to report whether the second word they heard was *tapes* or *capes*. Listeners reported hearing *capes* when the token was preceded by *Christmas* (i.e. hearing /k/ when preceded by /s/), and *tapes* when the token was preceded by *foolish* (i.e. hearing /t/ when preceded by /ʃ/) with consistent, significant accuracy.

Elman & McClelland (1988) concluded that these results were due to the perceived coarticulation between adjacent sounds, even though there was none. Top-down influences lead listeners to definitively perceive the same ambiguous fricative as either an /s/ at the end of *Christma_* or an /ʃ/ at the end of *fooli_*, as those sounds are the only possible candidates that make English words. Thus, participants' perception of the ambiguous stop as being either a /t/ or /k/ can be attributed to the perceived coarticulation between the adjacent sounds. The coarticulatory effect here can be explained in terms of relative frequency. The English phoneme /s/ has a relatively high frequency, and so listeners attributed any high-frequency noise they heard in the ambiguous stop as being due to carryover coarticulation from the contextually perceived /s/. Listeners therefore reported perceiving the low frequency phoneme /k/ out of the ambiguous stop token. Thus, participants reported hearing *Christmas capes*. When the context lead listeners to perceive an /ʃ/, which has relatively low frequency, they attributed any low-frequency noise in the ambiguous stop as being due to carryover coarticulation, and therefore reported hearing *foolish tapes*. Listeners compensated for the expected preservative coarticulation between the two adjacent tokens, and restored different phonemes from the same ambiguous sound, depending on the preceding context.

This results described above can be taken as evidence supporting the reality of the TRACE model of spoken word recognition. A strictly bottom-up model may not have categorized the ambiguous fricative and stop consonants as readily as a model that integrates lexical information, such as TRACE. The ambiguous sounds in Elman & McClelland (1988) were acoustically identical. Therefore, if context played no role in phoneme perception, we could expect participants to perform at chance, choosing *tapes* and *capes* with at-chance significance.

In opposition to top-down models such as TRACE, strictly bottom-up models suggest that listeners process speech in one direction as it unfolds, accessing meaning based on the individual pieces. Such bottom-up models do not propose to integrate top-down feedback from higher-level processes such as context or word meaning. An example of such a model is the Cohort model, proposed by Marslen-Wilson & Tyler (1980). According to Cohort, as spoken words unfold over time, listeners compile a mental list of possible lexical candidates, called *cohorts*. For example, if the listener hears /sp/, he will compile the cohorts *spring, speech, spark, spill,* among countless others, all of which are possible lexical entries that begin with the phoneme sequence /sp/. As the remainder of the word unfolds, cohorts are eliminated until the recognition point is reached. The recognition point is the time at which one cohort remains and is selected for further integration into the rest of the speech stream (which is often but not necessarily the end of the word). Importantly, this model is purely bottom-up, and therefore does not consider context when choosing a cohort. The goodness of fit or cloze probability of a

possible cohort into the semantics of the sentence does not influence this model's recognition point decision or choice of possible cohorts.

Evidence supporting the Cohort model of spoken word recognition comes from Allopenna, Magnuson, & Tanenhaus (1998). The researchers tracked participants' eye movements across a grid using the visual world paradigm. In the visual world paradigm, objects bearing specific qualities are placed around a grid. Participants hear spoken instructions regarding where to look or how to rearrange objects on the grid. In the Allopenna et al. (1998) experiment, objects on the grid fell in to one of four categories based on the phonological shape of their name. Each object was either the *referent*, the object meant to be moved by the participant; the *cohort*, an object with a name beginning with the same two phonemes as the *referent*; the *rhyme*, an object beginning with different phonemes, but rhyming with the *referent*; or the name of the object was *unrelated*, creating a baseline condition.

During the course of the experiment, participants were instructed to "Pick up the [*referent*]; now put it below the [*object name*]" (Allopenna et al., 1998). The name of the referent was presented auditorily using the gating method (Grosjean, 1980). In this paradigm, spoken words are presented to listeners in phoneme-sized segments, always starting at the initial phoneme and gradually increasing in size until the whole word is presented. The goal of this task is to identify the recognition point of a given word. Allopenna et al. (1980) were interested in the effects that the *cohort* versus the *rhyme* and *unrelated* objects would have on participants' eye movements. If the Cohort model is a cognitive reality, this would predict that there would be competition between the *referent*

and the *cohort*, causing participants to look at both equally, until the gating revealed the true referent beyond doubt.

Results supported this prediction. Competition was reflected in eye movements between only the *referent* and the *cohort* when words were presented. Gradually, as the gate revealed more phonemes, participants' eye movements toward the *cohort* dropped. The *rhyme* and *unrelated* objects were not significant distractions to listeners, even as the gate expanded. Thus, these results support a bottom-up model such as Cohort.

Having reviewed two regnant and contrasting models of spoken word recognition, there remain a few more explanations to consider. Similar to and building upon the Cohort model, Gow & McMurray (2007) proposed the *continuous acoustic integration hypothesis*. They postulated that listeners use subphonemic cues to their advantage, and that integrating them into speech perception can lead to earlier lexical disambiguation. Listeners making use of subphonemic cues can make more accurate predictions about what will unfold next, facilitating spoken word processing. By the same processes, however, incorrect forms may remain active for longer. Essentially, Gow & McMurray (2007) stated that both phonemic and subphonemic cues are processed indistinguishably during spoken word recognition. Coarticulatory variations, therefore, may be encoded into speakers' phonological representations. This method of storing lawful subphonemic variations can be likened to the mental storage that resembles allophonic variation.

An additional bottom-up model of spoken word recognition that will prove relevant to this topic is the Merge model (Norris, McQueen, & Cutler, 2000). The Merge model, unlike TRACE, does not necessitate getting feedback from the mental lexicon.

Merge is like unlike Cohort in that it allows the lateral inhibition between lexical competitors. Lexical decisions in Merge are only achieved after word recognition has been successful. In this way, Merge is a purely bottom-up model that has no inhibitory connections between nodes at the feature or phoneme level. Merge resolves coarticulation in the following way: the more extreme the variability is in the speech stream mandates the strength of the influence the model receives from the lexicon. Finally, phonemic and lexical information is merged only at the decision stage. The Merge model appears to be the best means of fitting coarticulation into a model of spoken word recognition.

1.5 Acoustic Characteristics

The results demonstrated by Archibald & Joanisse (2011) mixed a wide variety of unbalanced onset and nucleus tokens, upon which no conclusions can be drawn regarding the PMN's behaviour to certain consonants or vowels. Thus, future studies are warranted that can systematically draw comparisons between meaningful consonant and vowel interactions. This is the focus of the current study, which requires a review of the acoustic characteristics of relevant consonants and vowels. Several phoneticians have studied the acoustic characteristics of sound classes at length. Yeni-Komshian & Soli (1981) reported that coarticulation is not manifested with identical robustness across phoneme classes. Thus, modulations found in the PMN component reported in Archibald & Joanisse (2011) may be due to unknown interactions between their numerous sound class combinations.

1.5.1 Consonants

Consonant classes differ greatly with respects to the way in which they encode cues to surrounding segments. Wright (2004) gives an overview of perceptual cues associated with different sound classes and reviews their perceptual salience. The present study focuses on stop consonants, so they will be the focus of this review.

Stop consonants are produced with a complete obstruction of airflow at a constriction site (place of articulation), and in the case of oral stops, the obstruction is followed by a burst of noise as the air pressure is released from the constriction site. In the case of nasal stops, the constriction site remains completely obscured, the velum is lowered, and air flows out of the nostrils (Raphael, 2008). Though stop consonants are temporally brief and noisy (Stevens, 2002), they carry several cues that reveal the identity of an upcoming vowel.

The first relevant cue in Wright (2004) is the formant transitions. Formant transitions are present in the juncture between a consonant and vowel. In a formant transition, the consonant constriction affects adjacent vowels. This causes the vocal tract to become deformed, resulting in perturbations of the formant structure of the vowel segment. Formant transitions are a very strong cue to identify surrounding constrictions, and thus giving listeners a preview of the place of articulation of an upcoming consonant.

Stop release bursts can also carry information regarding the identity of upcoming vowels. This is of particular interest to the present study, as all coarticulatory incongruencies occur at the juncture between a stop consonant and a vowel. Cues to upcoming vowels could be found in stops as early as at the release of the stop (Feng, Hao, Xue & Max, 2011). Upon release, the air pressure that built up behind the consonant

obstruction point is released, resulting in a brief, high amplitude burst. Stop bursts are aperiodic noise, and last a mere 5-10 ms. Wright (2004) states that this cue is important, but reliable only in unnaturally silent environments. This type of environment is exactly the type used in the present experiment, so this cue is crucial. The F2 transition in the burst of the stop reliably carries the identity of the upcoming vowel.

Nasal consonants are a rich source of cue encoding for consonants, since they are highly salient. The F2 transition in nasals is sustained, and like stop consonants, is a reliable, powerful cue to vowel identity (Wright, 2004). Additionally, nasals present a weakening of the higher formants due to antiresonance, and a low frequency resonance, called a pole-zero pattern. Found in this pattern is another cue to vowel identity.

1.5.2 Vowels

In comparison to consonants, vowel production allows air to flow freely and constantly out of the oral cavity without constriction. Traditional reviews of vowel perception rely on the relative spacing between F0 and the resonances – F1, F2, and F3 (Reetz & Jongman, 2011). The patterning of relative formant frequencies reveals height, backness, and rounding of a vowel segment. Generally, the value of F1 is inversely related to vowel height, whereby raising the tongue decreases the value of F1. The value of F2 relates to vowel backness. Front vowels have a high F2 value, and the value of F2 decreases as the tongue approaches the back of the oral cavity. Thus, vowel identity is reliably extracted from the relative differences between F0, F1 and F2. Figure 1 below is an illustration of relative resonances.

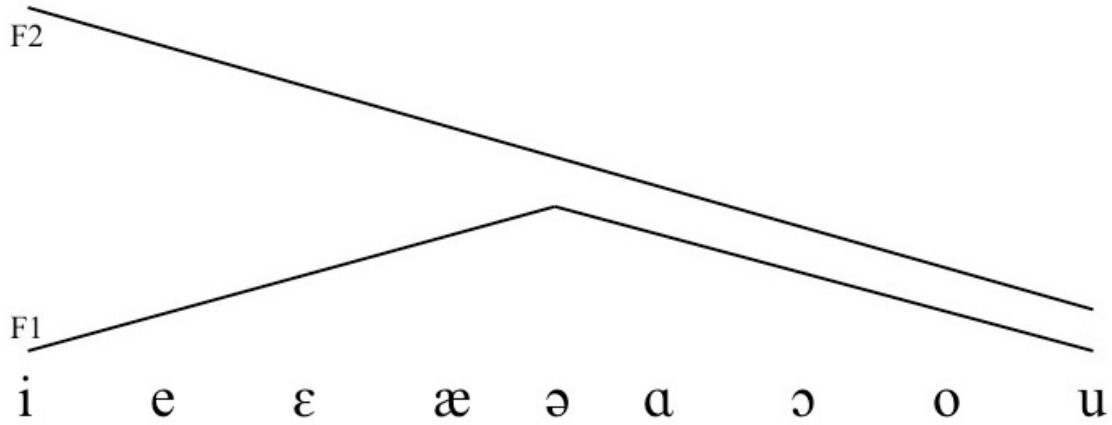


Figure 1 Figure illustrating relative differences between the first and second formant for each of the English vowels.

Identifying a vowel solely based upon the relative frequencies of F0 and its resonances is viably only true for carefully articulated vowels that reach and maintain an extended steady-state segment. In natural speech, vowels rarely ever achieve, let alone maintain a steady-state portion. Vowels that are flanked by consonants have formants that reliably fall short of their ideally articulated targets. Due to this articulatory "undershoot", vowel identity is more reliably extracted from the consonant-vowel juncture than from relative distance between steady-state formant frequencies (Wright, 2004).

2.0 Present Study

Evidence provided from the literature establishes subphonemic cue processing as a topic worthy of further investigation. As the results of Archibald & Joanisse (2011) and Arbour (2012) demonstrate, the PMN is sensitive to coarticulatory miscues. Thus, the present study employs the same ERP methodology in order to further investigate the PMN's sensitivity to the phenomenon of coarticulation.

Arbour (2012) addressed several issues inherent in the Archibald & Joanisse (2011) study. For example, one major limitation of Archibald & Joanisse (2011) was the wide and unbalanced range of consonant-vowel combinations. This variety was necessitated by their presentation constraint of only using imageable stimuli. Arbour (2012) used a written-word/spoken-word paradigm, drastically increasing the bank of possible word stimuli. Thus, Arbour (2012) was able to draw comparisons across consonant classes (i.e. fricative vs. stop onsets).

In her analysis, Arbour (2012) examined differences between classes of onset type, looking separately at the waveforms for words beginning with stops versus fricatives, and at the differences between high and low vowels. Significant differences in the PMN were revealed in analysis. The present study aims to complete a more complex and intricate analysis of vowels and consonants.

The present study limits onset consonant class to strictly stops – both oral and nasal. Stop consonants were limited in place of articulation to labial and alveolar, in order to maintain a balanced set (the velar nasal does not occur word-initially in English). Thus, onset consonants varied along three binary features: [+/- bilabial], [+/- oral], [+/- voice] for the 6 stops /p, b, m, t, d, n/. Vowels were limited to the four corner vowels of English (/i, u, æ, ɑ/), maximizing the distance across the vowel space (c.f. adaptive dispersion theory, Liljencrants & Lindblom, 1972). Extensive coding was imparted on the stimuli, specifying details of the consonant-vowel juncture of each stimulus. In this way, analyses could be completed to compare many consonant-vowel pairs, as well as

distances between vowels, to reveal whether the PMN is sensitive to such minute subphonemic details.

In keeping with Arbour (2012), stimuli were organized into minimal sets, to achieve maximally balanced stimuli. The target juncture was the transition between the onset consonant and nucleus of monosyllabic English words that had CVC structure. According to Wright (2004), word onset is a optimal place to study coarticulation, as segments in word-initial position undergo a phenomenon he calls "boost at onset". This phenomenon describes the increased salience of word-initial segments as speech processing is "kickstarted". Coda consonants were kept constant to minimize any regressive coarticulatory effects.

Consistent with the findings from Archibald & Joanisse (2011) and Arbour (2012), it was hypothesized that the PMN would be sensitive to differences in the consonant features described above. Since voiceless stops are aspirated in English, cues to the upcoming vowel can be found in the aspiration of each consonant. Aspiration noise is similar to the unobstructed fricative noise of an /h/ token, which was found in Archibald & Joanisse (2011) to produce the most robust PMN. Therefore, it is hypothesized that the PMN will have increased sensitivity to coarticulatory miscues following a voiceless stop as compared to voiced. Since nasal stops generally have a longer duration compared to oral stops, it is hypothesized that the PMN will be more sensitive to miscues following a nasal than oral stop. Finally, it is hypothesized that bilabial stops will interact significantly with vowel backness, as the distance between the stop and vowel in the oral cavity is greater than the distance between an alveolar stop and

any given vowel. As for vowels, it is hypothesized that distance across the oral cavity (i.e. high-back to low-front is greater articulatory distance than high-back to low-back) will modulate the PMN response.

3.0 Methods

3.1 Participants

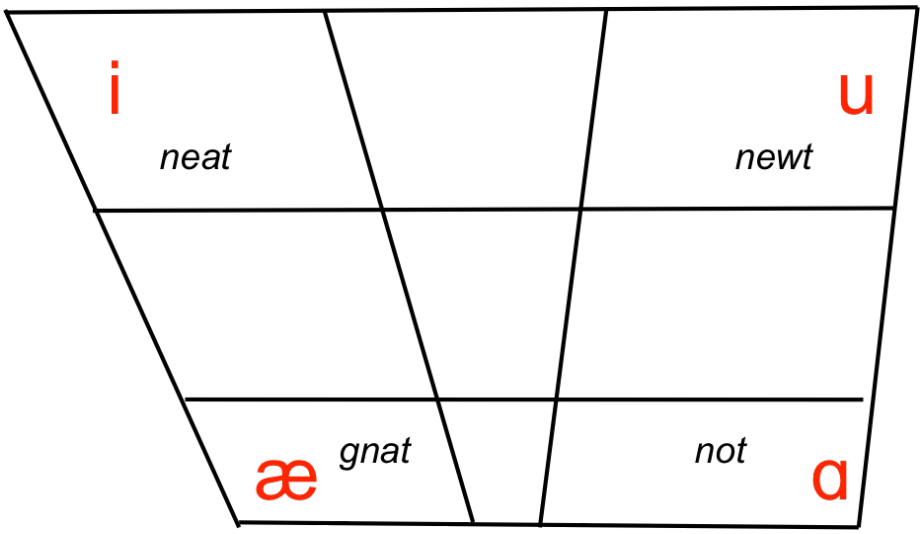
Twenty-two native English-speaking undergraduates (14 female), ranging in age from 18-36 years participated in the experiment through the department's course credit system enabling them to receive credit in exchange for their participation. All participants were typically developed, with no reported history of neurological or hearing impairments, normal or corrected-to-normal vision, and were not taking any medication at the time of testing. This study was approved by the McMaster Research Ethics Board and informed consent was obtained prior to testing. Letters of Information, Screening Forms, and Debriefing Letters can be found in Appendices 3, 5 and 5.

3.2 Stimuli and Experimental Conditions

Stimuli consisted of 76 monosyllabic CVC-structured English words. These words were organized into minimal sets within which words differed by only one phoneme (the vowel nucleus) while the onset and coda consonants were static. Keeping the coda constant eliminated any effect of differing regressive coarticulation. Each word was created using one of the four corner vowels of English /i, u, æ, ɑ/, which exploited the maximal distance across the vowel space. Each vowel was given an onset of an anterior stop /p, t, b, d, m, n/ and a coda of an oral stop /p, t, k, b, d, g/, provided that the

combination resulted in an English word. Table 1 below shows an example of a full set of stimuli.

Table 1 One full set of stimuli, organized onto the vowel space. Onset /n/ and coda /t/ consonants remained static. The nucleus (vowel) was interchanged between one of the four corner vowels of English. NB: speakers were not included in stimuli recording if they pronounced "newt" with a glide (/njut/).



Real words were rejected if their spoken frequency fell below 1 per million. Frequency data for all but two words were drawn from the SUBTLEX-US Corpus (Brysbaert & New, 2009). The SUBTLEX-US corpus collects frequency data from the subtitles of English language films, and is considered a viable representation of colloquial spoken English. The first of two exceptions, "toque", was not found in the corpus, but had a frequency of 1.04 in the Corpus of Global Web-Based English (GloWbE) (Davies, 2013) when the search was based on Canadian English. The second of the two exceptions, the word "Mac" did not meet the frequency threshold in SUBTLEX-US nor GloWbE but is a hypocoristic name for McMaster University. Since participants were

undergraduate students of McMaster, the word will have a very high frequency in their language usage. A full set of stimuli can be found in Appendix 1.

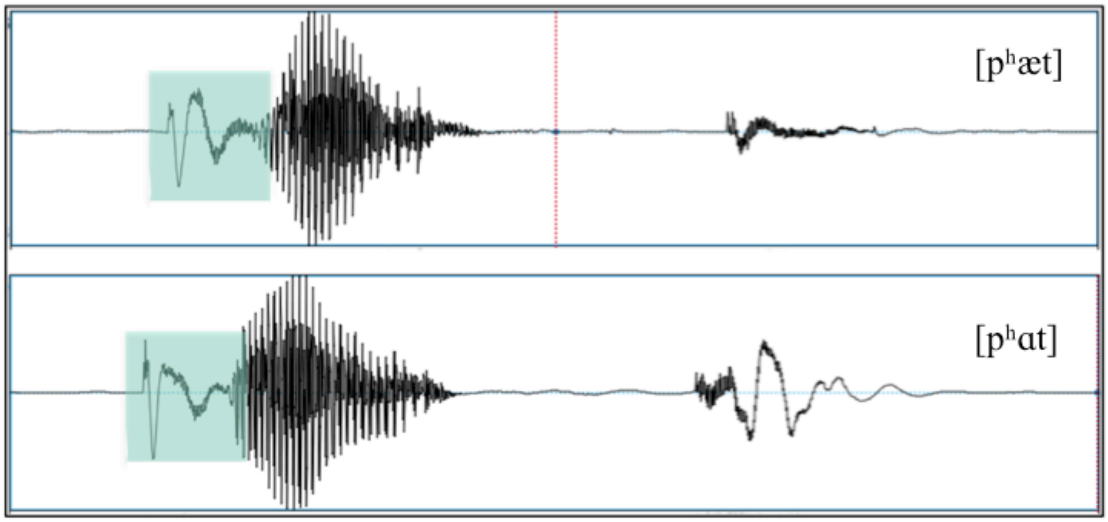
Following Arbour (2012), words were presented visually to participants rather than as pictures (Archibald & Joanisse, 2011). This method placed no restrictions on the "imageability" of the stimuli, thus widening the scope of possible stimulus tokens. Additionally, more focus could be placed on controlling the phonological environments of each stimulus, allowing for more detailed data analysis and comparison of phenomena. Expressly, vowels could be compared along two dimensions, and consonants could be compared within one natural class (i.e. within the class of stops, as opposed to comparing two classes, stops and fricatives). Written stimuli were presented to participants on a computer screen for 1500 ms. Immediately following visual presentation, words were played to participants through headphones. Words had a mean duration of 889 ms (SD=188 ms).

All stimuli were recorded from five female native speakers of Southern Ontario English. During the course of the experiment, participants heard 20% of the tokens produced by each speaker. The voice that produced each word was randomized between participants. This number of speakers and presentation method was chosen so that any experimental effects could not be attributed to an individual speaker's idiosyncrasies. Each word was spoken three times during recording sessions. To minimize coarticulatory effects across word boundaries, speakers were asked to bring their lips to a fully closed position between each production (Archibald & Joanisse, 2011). Speech tokens were recorded in Praat (Version 5.3.56) using an Audio-Technica ATM73A head-mounted

microphone, attached to a desktop computer using a TASCAM US-122MKII USB 2.0 Audio/MIDI interface. Speech was recorded at a sampling rate of 44.1 kHz. Stimulus amplitude was controlled by leveling all files to 75 dB using a custom Praat script.

The optimal spoken word version was selected from the three produced by each speaker. Optimality was determined based on visual inspection of the waveforms regarding continuity of the pitch contour and an absence of creaky phonation, which could affect perception of the vowel. Optimal productions were selected such that each word on the list (see Appendix 1) was selected. Each file's waveform was viewed in Praat and the waveform was magnified to show the transition period from initial consonant to vowel. A zero-crossing point was identified at the point at which the consonant transitioned into the vowel of each word. A boundary line was placed during the transition, and function native to Praat moved it to the closest zero-crossing, which now marked a section of each file containing only the onset of each word. A custom Praat script was designed to splice onsets from one word to another of user-identified pairs. Congruent or identity-spliced tokens were created by splicing the onset with a different production of the same word (different from the one selected to appear in full in the experiment). For example, splicing a production of mood /m^{w}ud/ with another production of mood creates the congruent token [m^{w}ud]. Incongruent or cross-spliced tokens were created by splicing a word with another from the same minimal set. Splicing a production of mood /m^{w}ud/ with a production of mead /m^{j}id/, creates the incongruent tokens [m^{j}ud] and [m^{w}id] (McQueen, et al, 1999; Archibald & Joanisse, 2011). Figure 2 shows waveforms that illustrate the splicing process.

Figure 2 Waveforms showing congruent coarticulation of <pat> and <pot>. The shaded area shows the onsets that were cut and cross-spliced to create incongruent coarticulatory information at the CV-juncture.



Three experimental conditions were created with this splicing method (Table 2). In the **Congruent** condition, participants were presented with an auditory token that was lexically congruent as well as having congruent coarticulatory cues throughout. In the **Incongruent** condition, the auditory token was a lexical match but had incongruent coarticulatory cues at the CV-juncture. The third condition (**Unrelated**) used coarticulatory cues that were congruent, but the spoken word was lexically incongruent to the written word. This last condition was included in the experiment to elicit a mismatch response that was obvious to participants, since the incongruity of the incongruent condition was difficult to detect.

Table 2 Table showing the three types of experimental conditions in the present study, with examples from each condition.

	Written Word	**Spoken Word**
Congruent		
Coarticulatory match *Lexical Match*	mood	[m^{w}ud]
Incongruent		
Coarticulatory mismatch *Lexical Match*	mood	[m^{j}ud]
Unrelated		
Coarticulatory match *Lexical mismatch*	food	[k^{j}ip]

Coding was applied to the stimuli in order to fully analyze types of incongruities after data had been collected. Each trial was given a three digit code that could then be decoded in a way that revealed the identity of the consonant, the original vowel with which that consonant was articulated, and the identity of the post-splicing vowel. For example, the word *mood* in the congruent condition carried the code 944 (9: [m], 4: [u] in the original articulation, 4: [u] in the post-splice environment). The incongruent token [m^{j}ud] carried the code 924. The second digit, indicating the original vowel, represents the [i] with which the [m] was originally articulated. The goal of this coding was that conditions could be closely compared in terms of phonological similarity or dissimilarity. Consonant conditions in analysis will be referred to by the feature being examined (i.e., voicing, place of articulation). Vowel conditions will be referred to in terms of *original articulation – post-splice vowel* (i.e., i-u, indicating that the onset contained cues to the vowel [i], but the nucleus in the token was the vowel [u]).

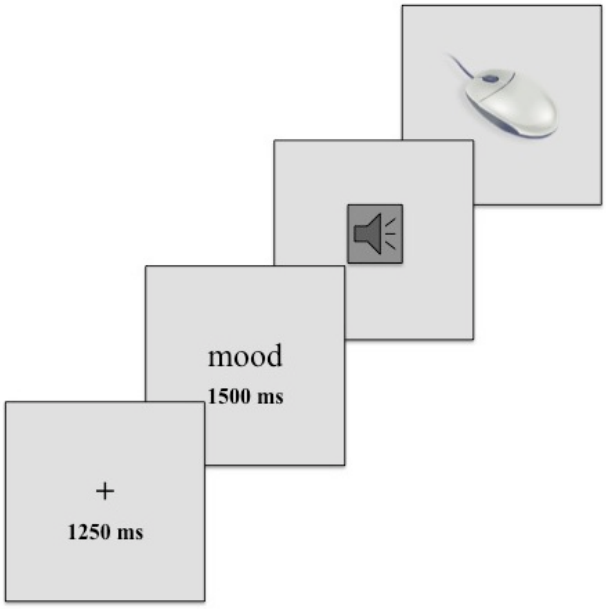
One testing period consisted of two successive sessions, each presenting participants with 315 tokens for a total of 630 tokens. Each unique sound file was presented once per session. Speaker was randomized so that each voice played 63 times per session, with no two speakers producing the same stimulus. The breakdown of trials per participant was as follows: 184 congruent tokens, 396 incongruent tokens, and 56 unrelated. Stimulus presentation was randomized across participants. Of particular interest during analysis is the comparison between pre- and post-splice vowel height and backness. Previous work (Archibald & Joanisse, 2011; Arbour, 2012) was unable to analyze effects of vowel type with this level of specificity. Codes were entered for each stimulus specifying both height and backness of the original vowel, height and backness of the post-splice vowel, as well as consonant type. Consonant type varied along three dimensions: oral/nasal, voiced/voiceless, and bilabial/alveolar. Participants were given five practice trials before each session, which were not included in the analysis. See Appendix 2 for a full list of trial counts and stimuli conditions.

3.3 Procedure

The experiment was programmed and presented using Presentation software (NeuroBehaviouralSystems Presentation 14.7). Participants were seated about 1m from a 24-inch Hewlett Packard LCD computer monitor. Each trial began with the presentation of a fixation cross for 1250 ms followed immediately by the presentation for 1500 ms of a single written word on the screen (white text on black background). This single word acted as the priming stimulus that created an expectation for which word participants were about to hear. A spoken word was then played to participants binaurally through

insert headphones (Etymotic Research) attached to an amplifier (ARTcessories HeadAmp4). Participants were instructed to indicate whether the word they heard matched the word that they had read. Responses were made on a computer mouse. A left button click indicated a match, and a right button click indicated a mismatch. The next trial began with a fixation cross after the participant had made their button click response. See Figure 3 for a depiction of one trial.

Figure 3 Illustration of the progression of a single trial using the written-word/spoken-word paradigm.

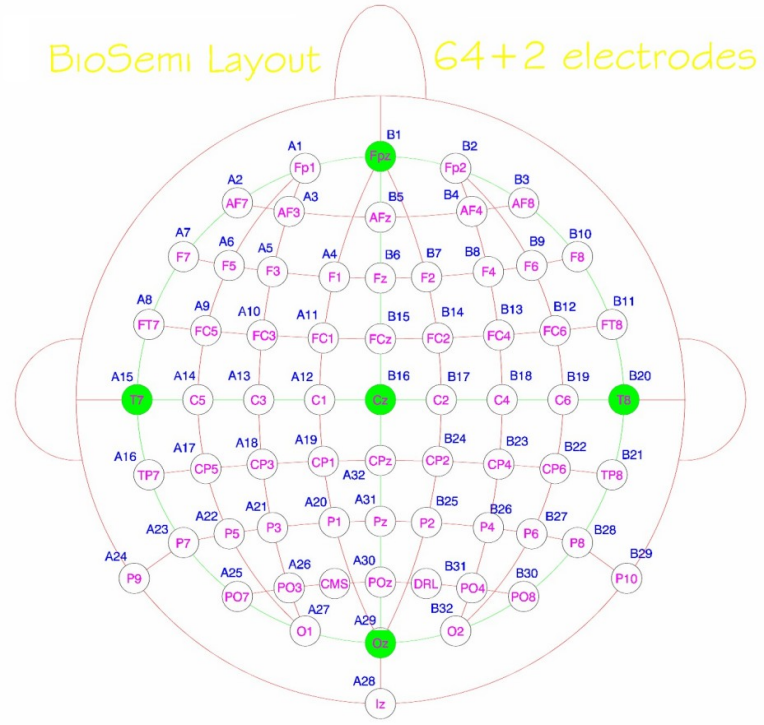


3.4 Electrophysiological Recording

Continuous electroencephalography was recorded using the BioSemi ActiveTwo system. Pin-type Ag/AgCl electrodes were used to record from 64 sites on a head cap, labelled according to the International 10-20 system (see Figure 4). EEG was recorded

using a bandpass of 0.01 Hz to 100 Hz, and sampled at 512 Hz. Online, data was referenced to the nose tip, and offline, data was referenced to the mastoids.

Figure 4 Layout of 64-channel setup used to record EEG from participants. Image retrieved from http://www.biosemi.com/pics/cap_64_layout_medium.jpg.



Continuous EEG was recorded throughout the duration of the experiment, but triggers were placed at specific times within each stimulus, marking relevant events. Markers were placed at the onset of the word, at the end of the consonant, including original coarticulatory information, and again at the beginning of the post-splice vowel. Electrooculographic (EOG) activity was recorded via external flat electrodes placed supraorbitally and on the outer canthus of the left eye. Information recorded at these sites

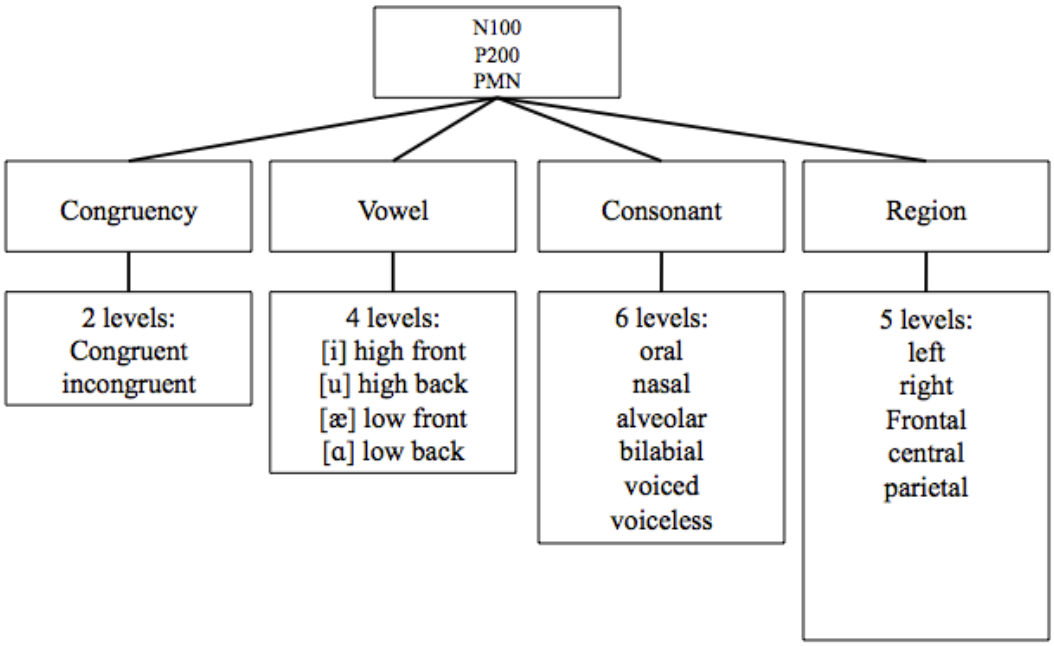
was used to remove ocular artifact due to horizontal and vertical eye movements and blinks from the EEG data.

Three ERP components were of particular interest to the current analysis: The N100, identified as the most negative peak between 80 ms -200 ms post stimulus onset; the P200, identified as the most positive peak between 200 ms – 300 ms; and, the PMN, identified as the most negative peak following the P200 between 250 ms - 350 ms. Time windows for these components were based upon previous literature (Newman & Connolly, 2009; Archibald & Joanisse, 2011; Arbour, 2012), and upon visual inspection of the waveforms.

Data were processed using Brain Vision Analyzer 2.0 (Version 2.0.4, Brain Products). Data were digitally filtered offline to a bandpass of 0.1-30 Hz. Waveforms were segmented and grand averages were created using native functions included in the program. Epochs for three ERPs were identified: N100 (80 ms – 200 ms), P200 (200 ms – 300 ms) and the PMN (260 ms – 360 ms). Brain Vision Analyzer's semi-automated peak finder function was used to identify the most negative/positive peak during each epoch. Data were then exported from Brain Vision Analyzer and imported into R (Version 0.98.994) for data analysis.

A repeated-measures analysis of variance (ANOVA) was conducted for each of the three ERP components identified above. Figure 5 outlines the factors and their levels in this study.

Figure 5 Flowchart showing levels for all ANOVA analyses for each ERP component.



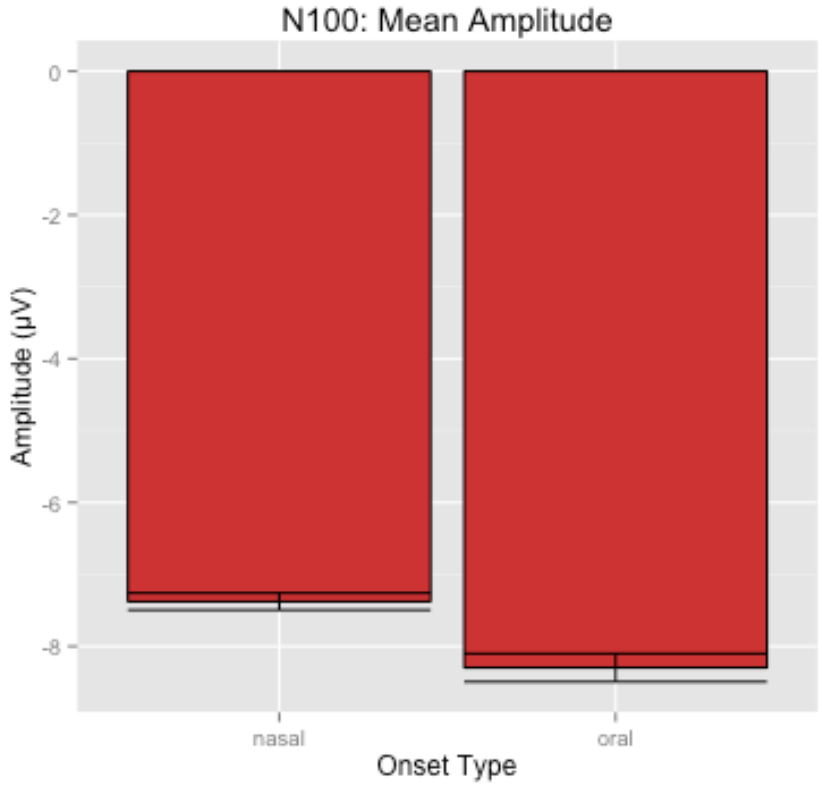
4.0 Results

4.1 N100

The amplitude of the N100 response was significantly modulated by various acoustic properties of word onsets. Phonologically, it is logical that these onsets be compared not individually, but by natural class parings. Thus, onsets were organized into phonological classes and analyzed as follows: oral vs. nasal, voiced vs. voiceless, and bilabial vs. alveolar.

Oral:Nasal - The mean amplitude of the N100 response was significantly larger for oral stops (M = -8.30 μV, SD = 0.02) than for nasal stops (M = -7.38 μV, SD = 0.77), $t(1) = 7.88$, $p < .001$.

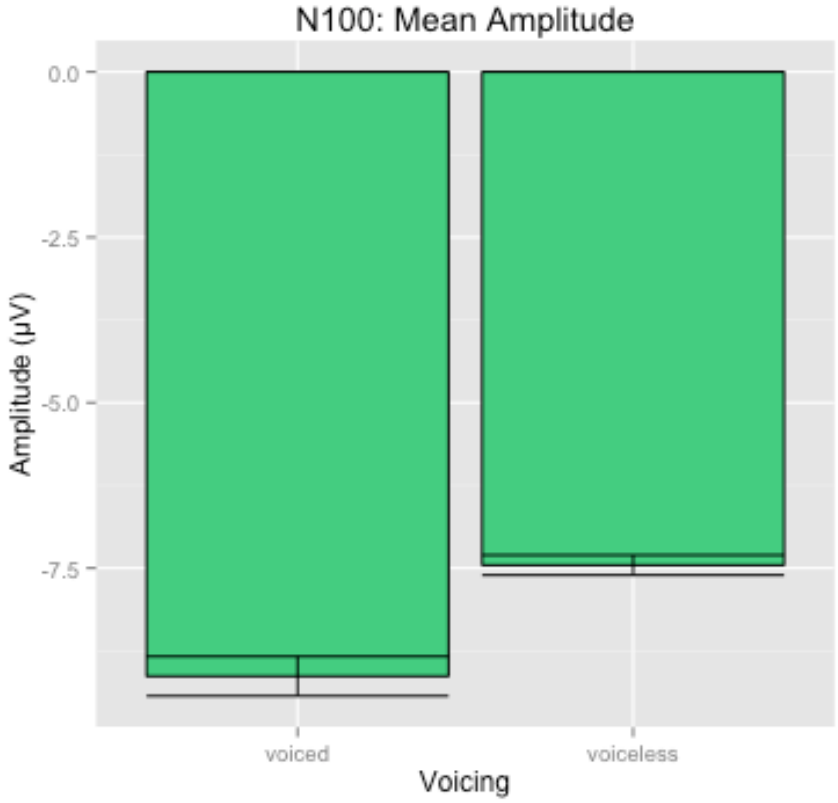
Figure 6 Mean amplitude of the N100 response for oral vs. nasal onsets.



Voiced:Voiceless - The mean amplitude of the N100 response was significantly larger for voiced onsets (M = -9.13 μV, SD = 0.02) than for voiceless onsets (M = -7.74 μV, SD = 0.01),

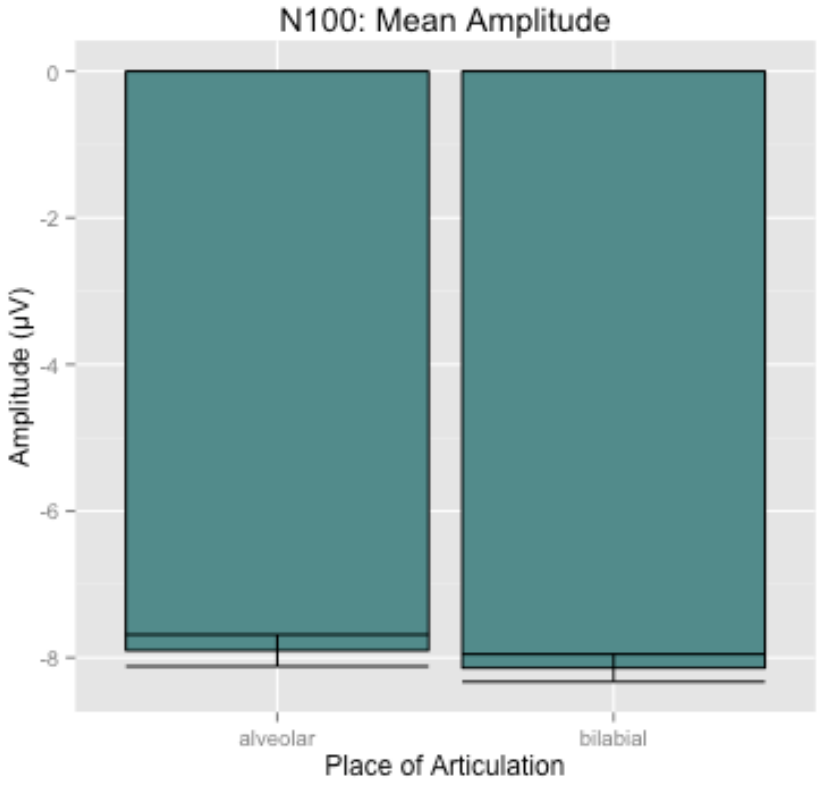
$t(1) = -9.83$, $p < .001$.

Figure 7 Mean amplitude of the N100 response by voicing status of onset (includes only oral stop consonants).



Bilabial:Alveolar - The third and final onset class pairing is place of articulation. The amplitude of the N100 response was not significantly modulated by place of articulation, $t(1) = 1.63$, $p = .103$.

Figure 8 Mean amplitude of N100 responses across place of articulation.



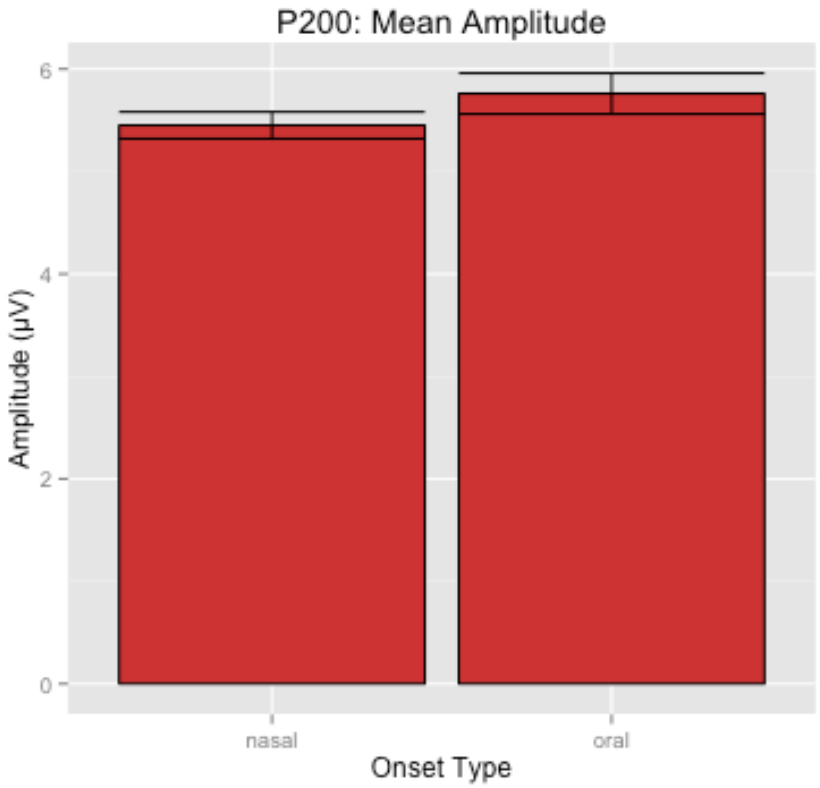
4.2 P200

Similarly to the N100 response, the P200 is plausibly only relevant to consonants in the onset position. The P200 was differentially elicited to onset type in Arbour (2012), showing a larger amplitude to stop type consonants, which is likely a reflection of the increased noise burst of aspiration in English. The P200 is not attributed to speech processing, but only to auditory stimulus detection (Martin, Tremblay & Stapells, 2007). Thus, onsets were organized identically to the previous section, and used to analyze the P200.

A *t*-test conducted on the mean amplitude differences between congruent and incongruent conditions revealed a significant difference, $t(1) = 3.05$, $p < .01$, where the mean amplitude of incongruent trials was larger than for congruent trials. An ANOVA revealed a significant main effect of onset type, $F(1,52) = 10.10$, $p < .01$. Thus, the P200 was analyzed in regards to onset consonant classes.

Oral:Nasal - The mean amplitude of the P200 response to oral as compared to nasal onsets was not significant, $t(1) = -2.55$, $p = .011$.

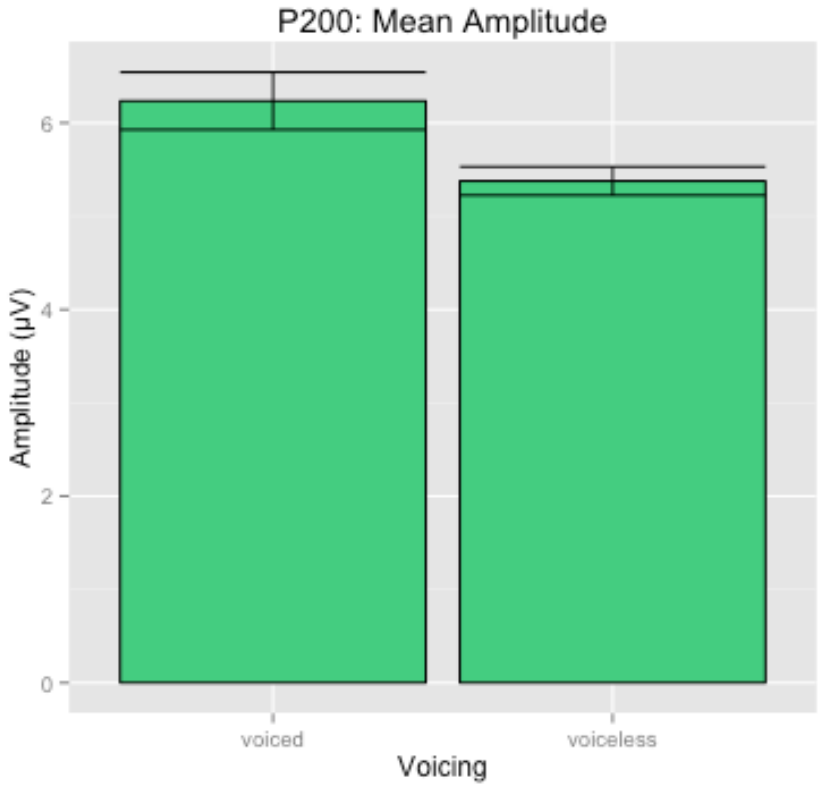
Figure 9 Mean amplitude of P200 response of nasal versus oral onsets.



Voiced:Voiceless - The P200 response was significantly modulated by voicing status of the initial consonant, $t(1) = 4.89$, $p < .001$. Nasals [m] and [n] were excluded from this comparison, since they have no voiceless counterparts in this study. The P200

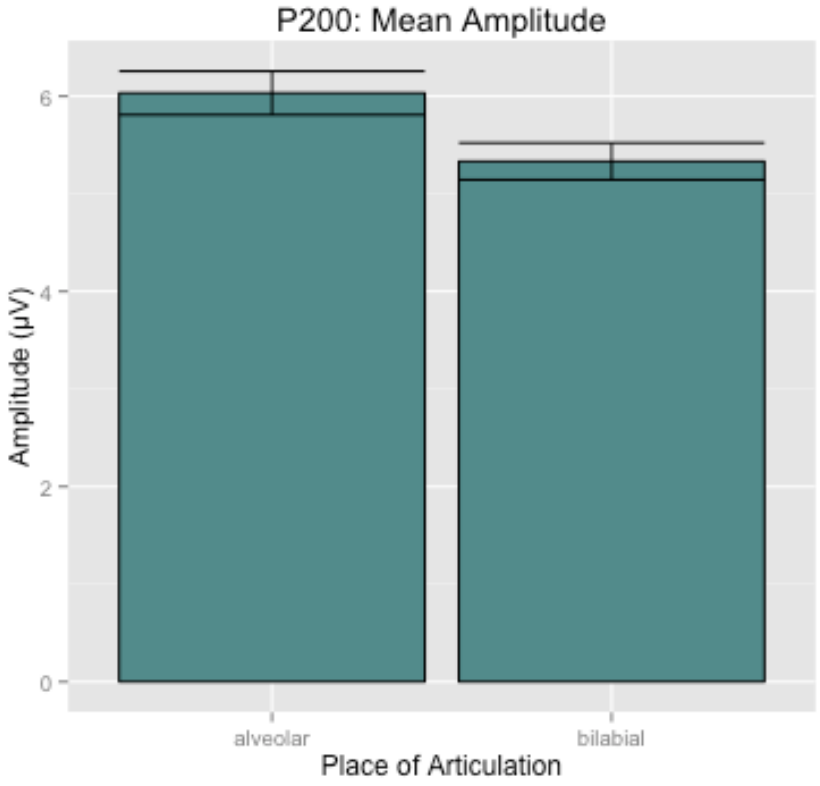
amplitude was significantly larger for voiced onsets (M = 6.24 μV, SD = 0.22) than for voiceless onsets (M = 5.38 μV, SD = 0.15).

Figure 10 Mean amplitude of P200 response for voiced onsets as compared to voiceless (only oral stops are included).



Bilabial:Alveolar - Finally, the mean amplitude of the P200 response was significantly larger for alveolar stops (M = 6.03 μV, SD = 0.18) than for bilabial onsets (M = 5.33 μV, SD = 0.16), $t(1) = 4.69$, $p < .001$.

Figure 11 Mean amplitude of the P200 response for alveolar vs. bilabial onsets.



4.3 PMN

The PMN has been shown to be sensitive to subphonemic cue incongruity (Archibald & Joanisse, 2011; Arbour, 2012) and has been proposed a marker of early, prelexical speech processing (Connolly & Phillips, 1994). The current experiment used carefully controlled and balanced stimuli to reveal the extent of the PMN's sensitivity to the phonetic properties of subphonemic incongruities. A repeated-measures analysis of variance (ANOVA) revealed three significant main effects. First, the ANOVA revealed a significant main effect of subphonemic congruity (congruent being less negative than incongruent), $F(1,47) = 48.10$, $p < .001$. The ANOVA also revealed a significant main

effect of caudality (most negative region being frontal), $F(2,3) = 33.17$, $p < .001$, and a significant effect of hemisphere (greater negativity found on the right hemisphere), $F(2,5) = 4.23$, $p < .05$. The analysis also revealed a significant two-way interaction between hemisphere * caudality, $F(6,11) = 3.20$, $p < .01$.

Figure 12 Grand average waveforms for all participants showing neural responses to congruent versus incongruent coarticulation.

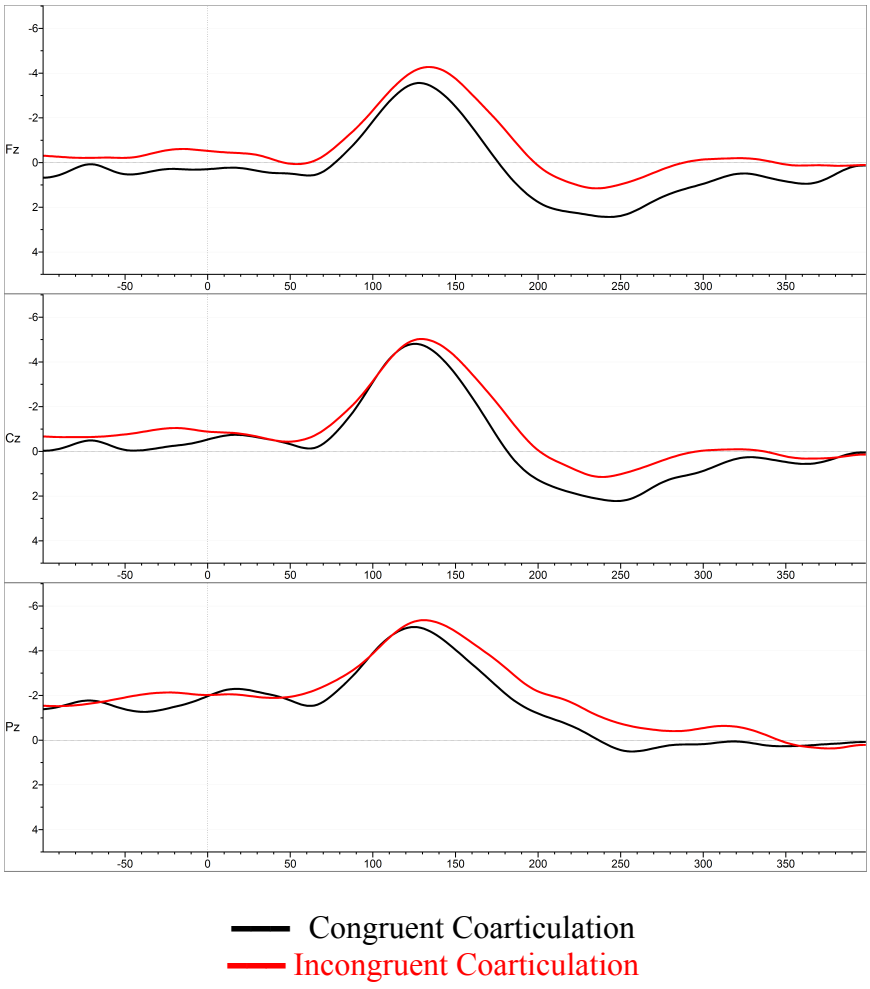
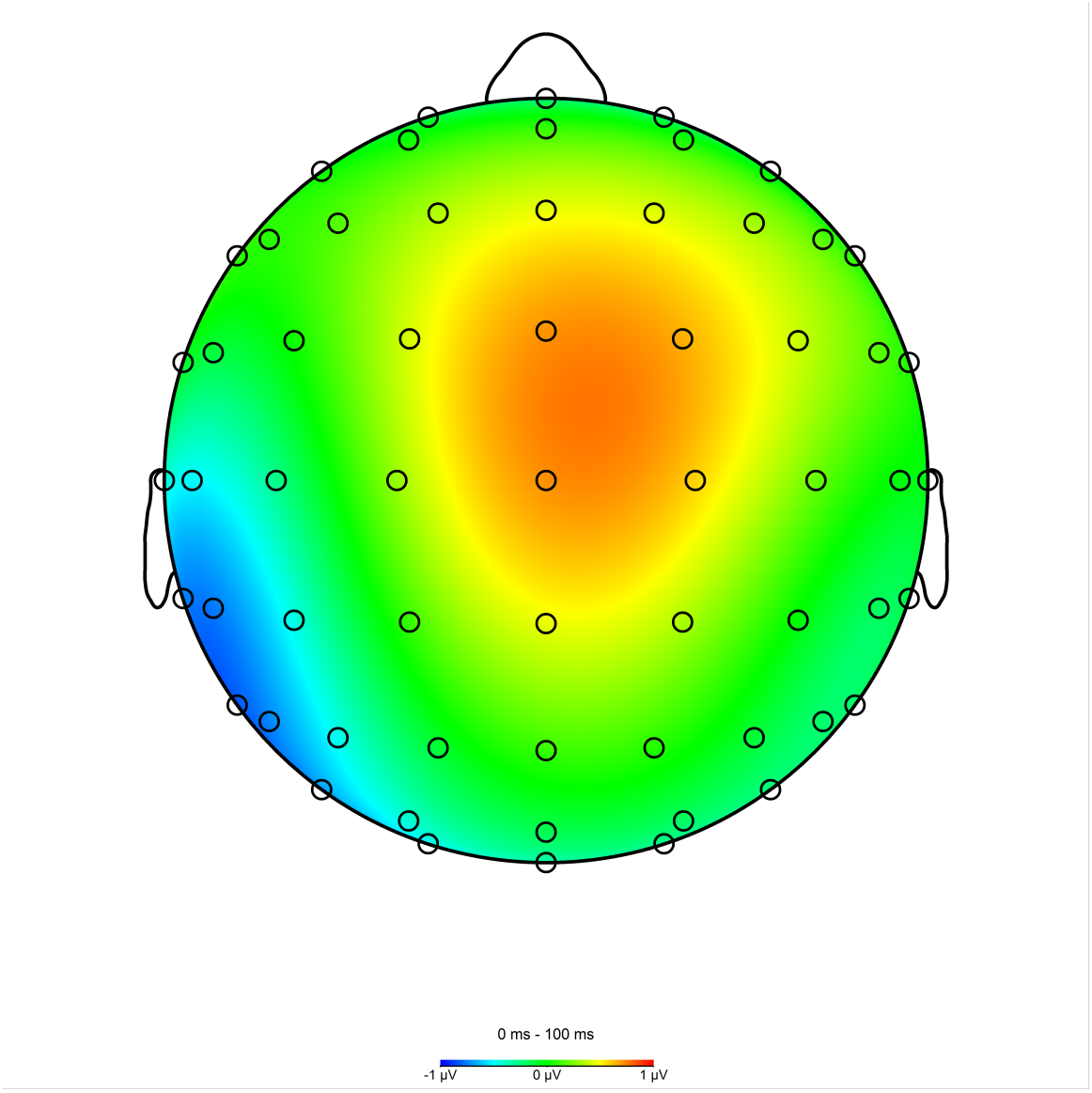


Figure 13 Topographical head map representing 64-channel recording of the PMN epoch, 260-360ms. This shows the effects of hemisphere*caudality described above, and shows that the PMN showed its typical scalp distribution of fronto-central.



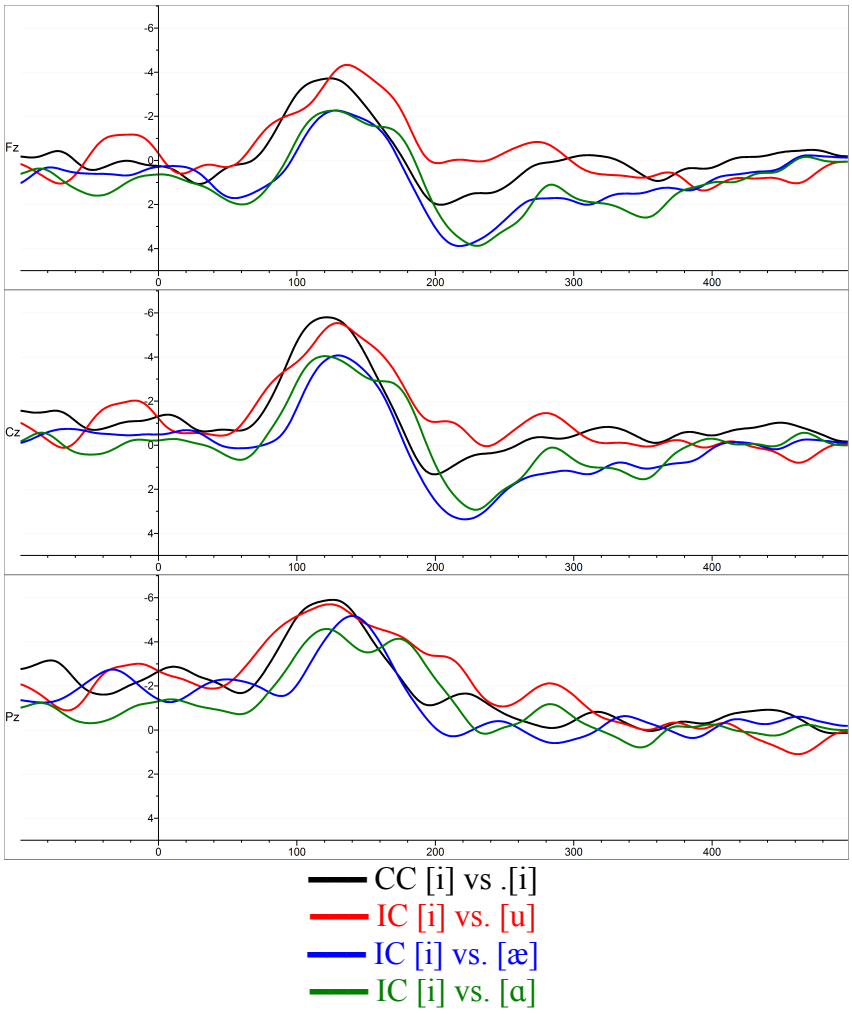
4.3.1 Vowels

The PMN data were divided into four groups based on the coarticulatory properties of each onset. Thus, cues to the "original vowel" are present in each onset. Incongruent conditions were those in which the nucleus in the target word was different

from the vowel whose cues were present at word onset. Since the four corner vowels in English were used as word nuclei, the method of splicing and balancing created one congruent condition to compare with three incongruent conditions (see Connolly & Phillips, 1994, for similar experimental format).

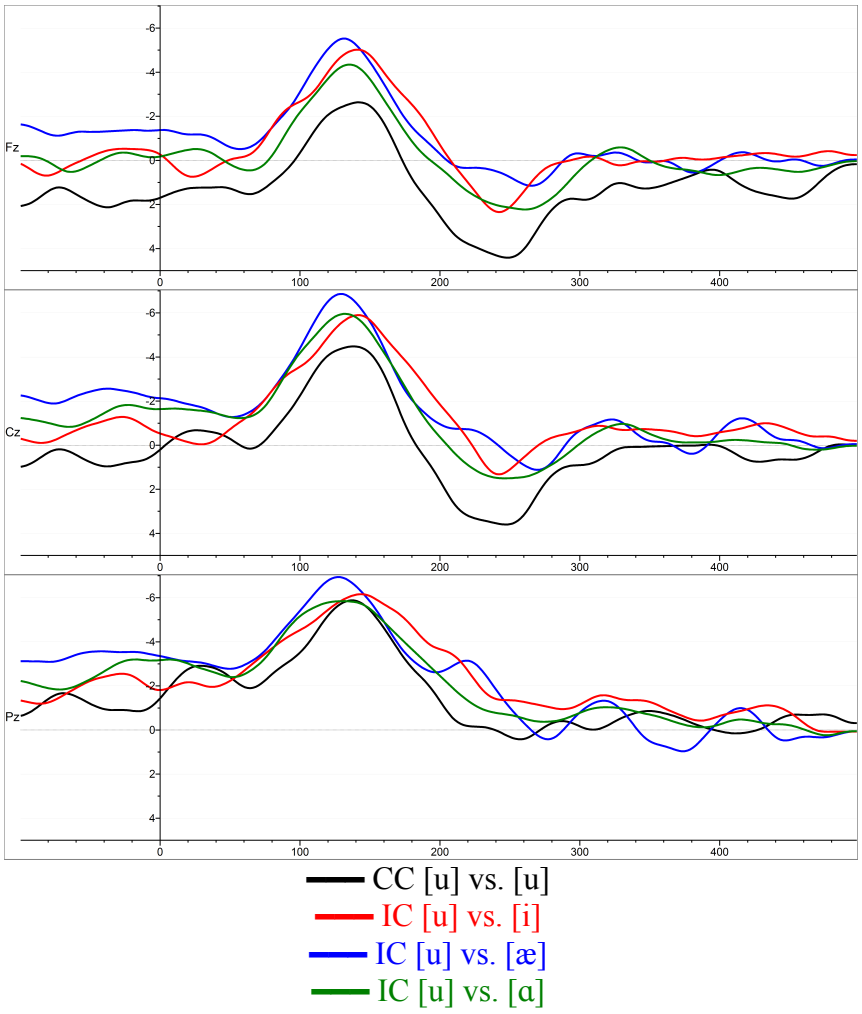
vowel /i/ – For words in which the subphonemic cues identified the vowel /i/, an ANOVA revealed a significant main effect of condition, $F(3,51) = 26.03$, $p < .001$. A pair-wise *t*-test was conducted to reveal significance between conditions. Results showed that the amplitude of PMN response in the congruent condition i-i was not significantly different from all three incongruent conditions. The PMN amplitude for condition i-æ was significantly smaller than those for i-u and i-ɑ, $p < .001$. However, the latter were not significantly different from each other.

Figure 14 Waveforms for all conditions in which the onset consonant carried coarticulatory cues to the vowel /i/ at sties Fz, Cz, and Pz.



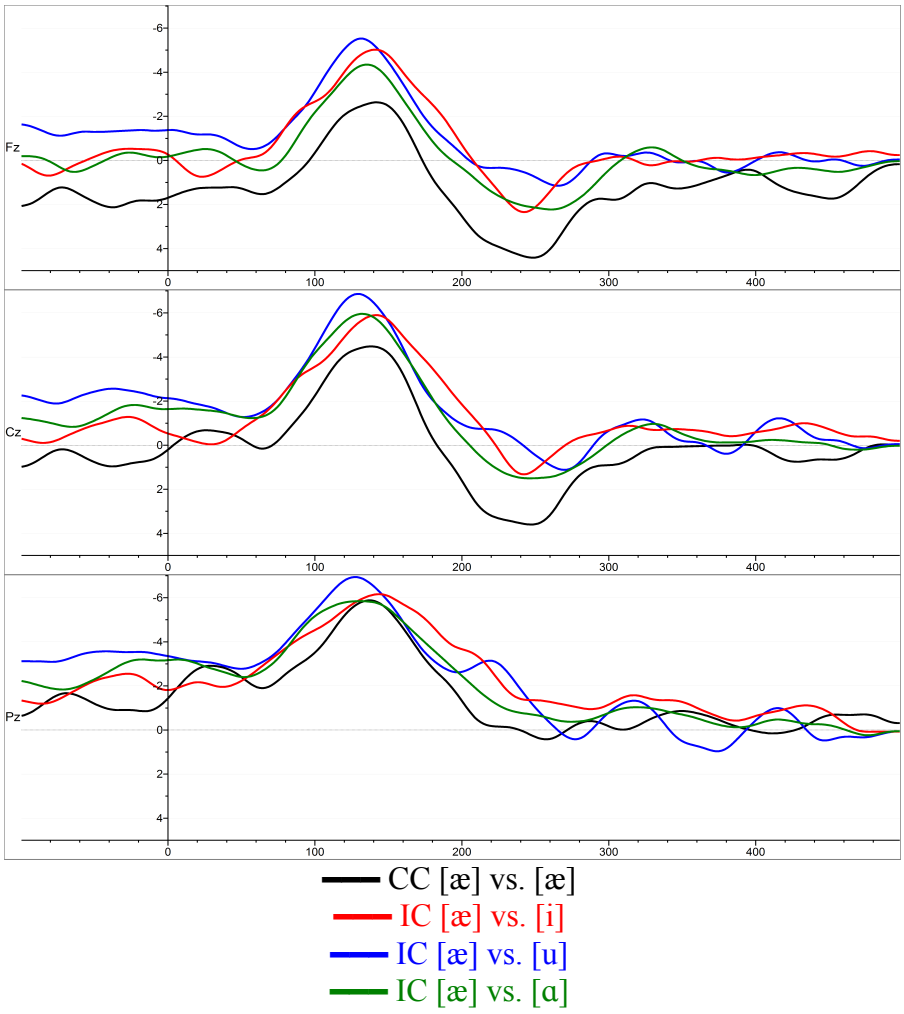
vowel /u/ – An ANOVA test conducted on all conditions in which coarticulatory cues in the onset were those of the vowel /u/ revealed a significant main effect of condition, $F(3,51) = 11.58$, $p < .001$. A pair-wise *t*-test revealed that mean PMN amplitudes between the congruent condition u-u and incongruent conditions u-i and u-ɑ were significant, $p < .05$. Condition u-æ was not significantly different from the congruent condition. The condition u-i was significantly different from both u-æ and u-ɑ, $p < .05$, but the latter two were not significantly different from each other, $p > .05$.

Figure 15 Waveforms showing neural responses to all conditions in which the onset consonant carried coarticulatory cues to the vowel /u/ at sites Fz, Cz and Pz.



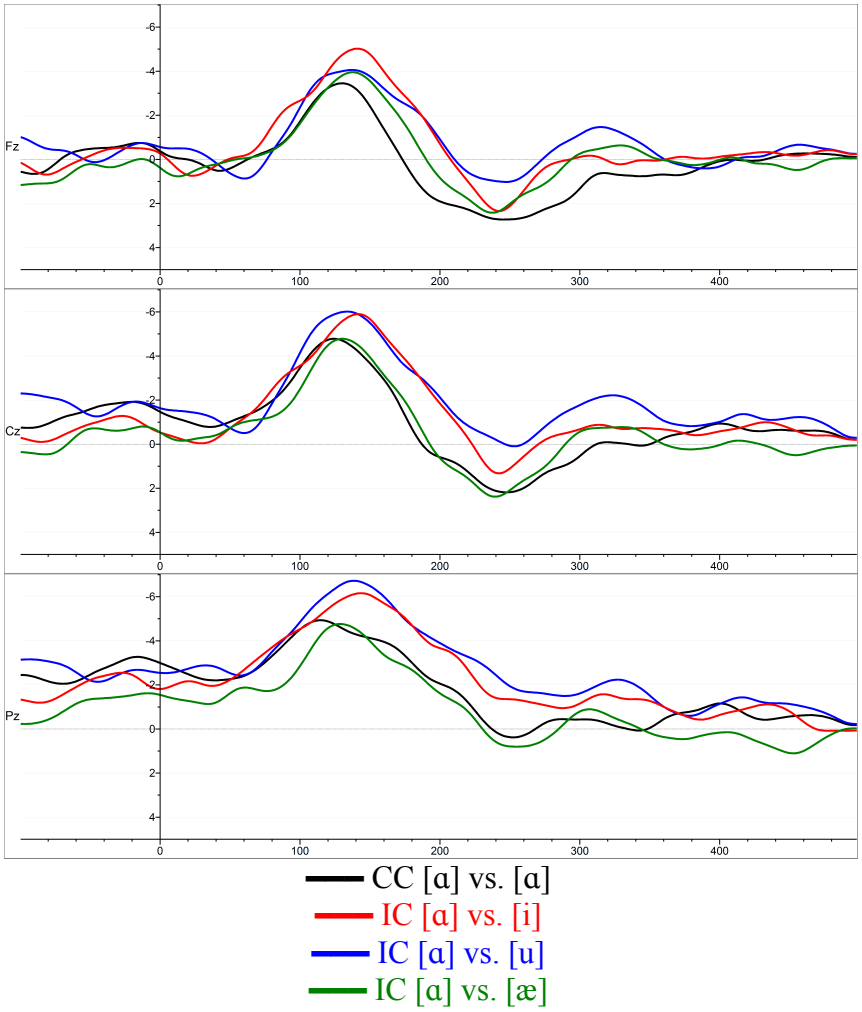
vowel /æ/ – An ANOVA conducted on all conditions where the onset consonant carried the coarticulatory cues of the vowel /æ/ also showed a significant main effect of condition, $F(3,51) = 113.2$, $p < .001$. A pair-wise *t*-test revealed that mean PMN amplitude between all incongruent conditions and the congruent condition were significantly different, $p < .05$.

Figure 16 Grand average waveforms for all conditions in which the onset consonant carried coarticulatory cues to the vowel /æ/ at sites Fz, Cz, and Pz.



vowel /ɑ/ – Finally, an ANOVA conducted on all conditions during which the onset carried cues to the vowel /ɑ/ revealed a significant main effect of condition, $F(3,51) = 50.70$, $p < .001$. A pair-wise *t*-test revealed that PMN mean amplitude was significantly different between each condition, $p < .05$.

Figure 17a Grand average waveforms for all conditions in which the original consonant carried the coarticulatory information of the vowel /a/, at sites Fz, Cz, and Pz.



Difference waves show the PMN more clearly. Figures 17b-d show the difference waveforms for all incongruent conditions where the original vowel was /ɑ/. Colours match the legend above. There is a prominent negativity visible in the waveforms of Figures 17b and 17c, both of which show difference waves for incongruent conditions that were significantly different from the congruent condition. Figure 17d, however,

displays a great deal of variance, with no prominent PMN, and depicts the condition that did not reach statistical significance.

Figure 17b Difference waveform for the incongruent condition a-i, for sites Fz, Cz, and Pz.

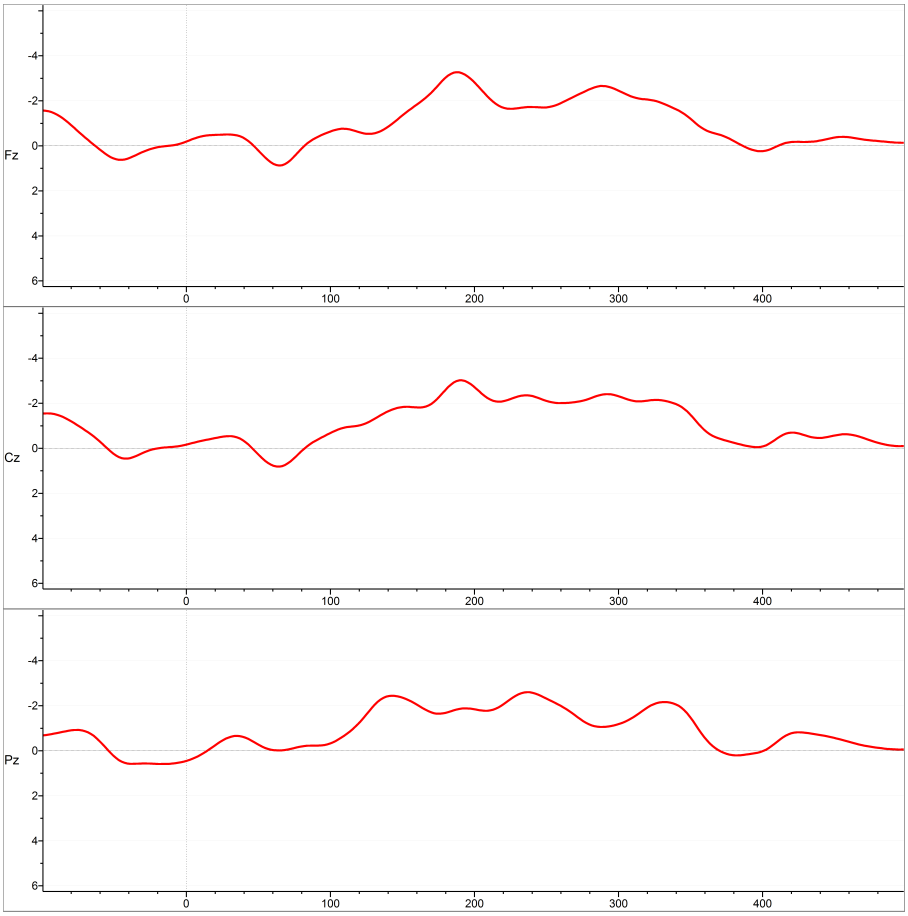


Figure 17c Difference waveform for the incongruent condition a-u, for sites Fz, Cz, and Pz.

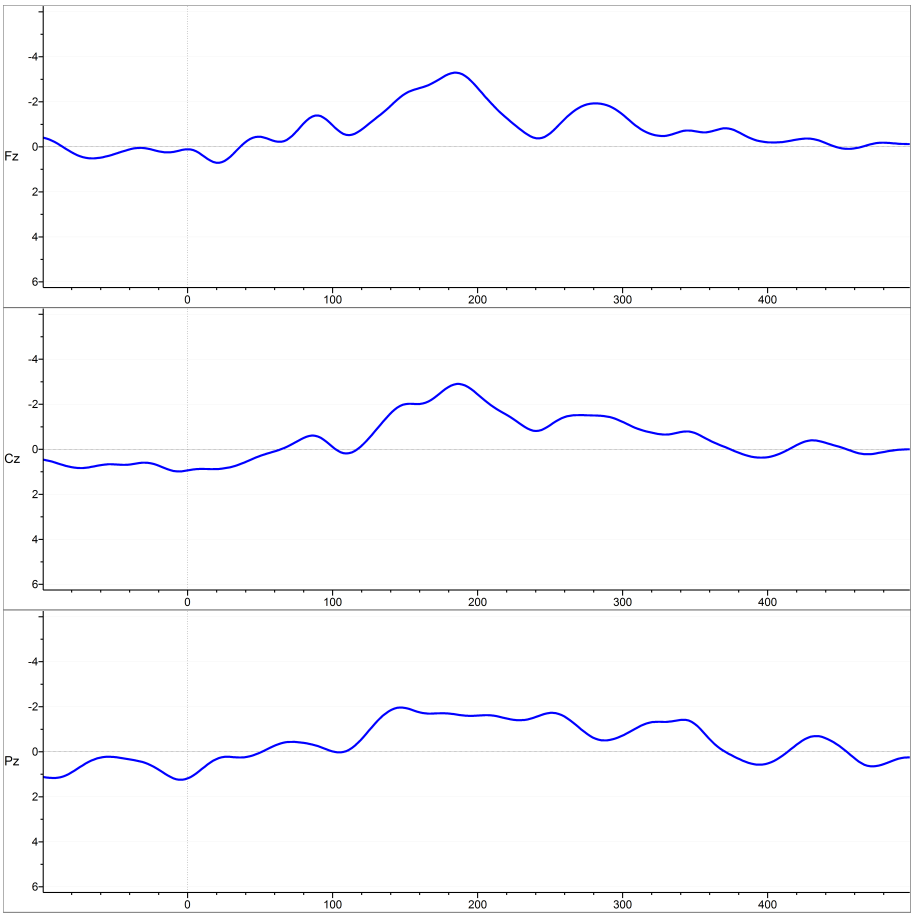
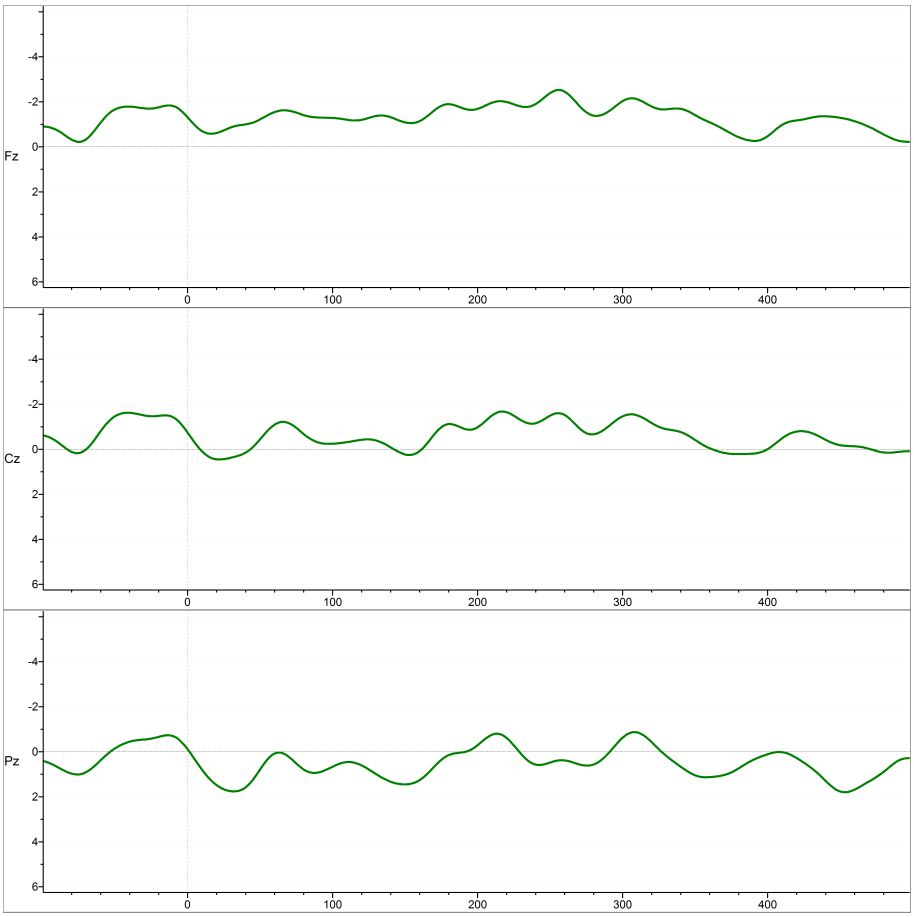


Figure 17d Difference waveform for the incongruent condition a-æ, for sites Fz, Cz, and Pz.

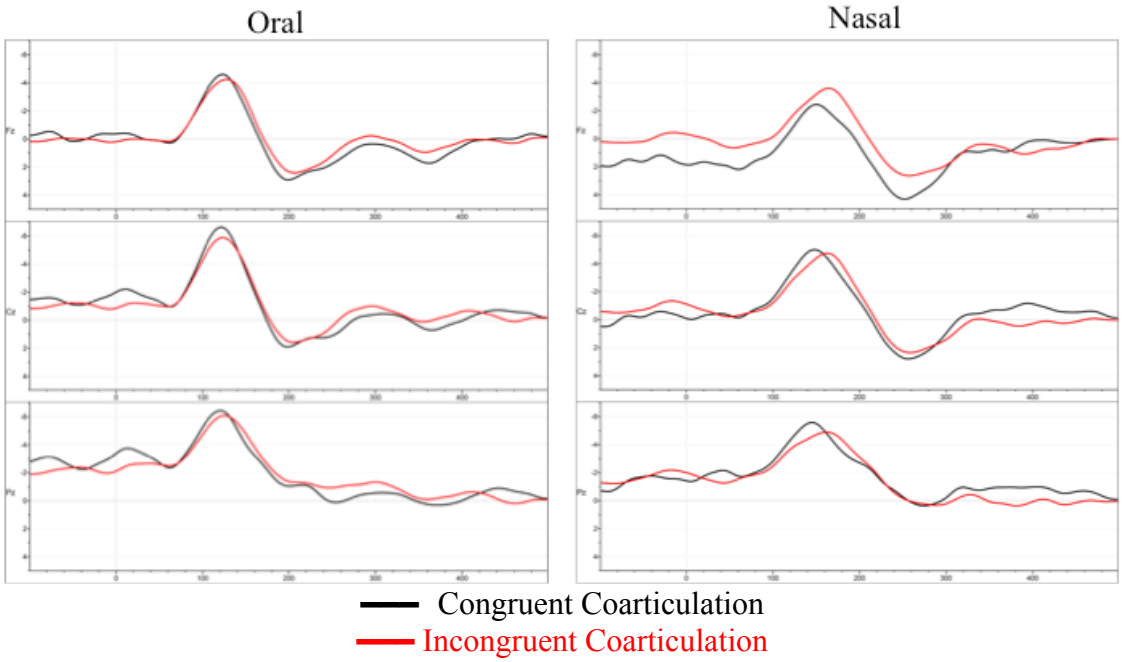


4.3.2 Consonants

The effects of consonant features on the PMN was investigated due to its significant main effect in the initial ANOVA test. Consonants were again grouped into natural phonological classes and compared with their balanced counterpart (i.e., oral/nasal, bilabial/alveolar, and voiced/voiceless).

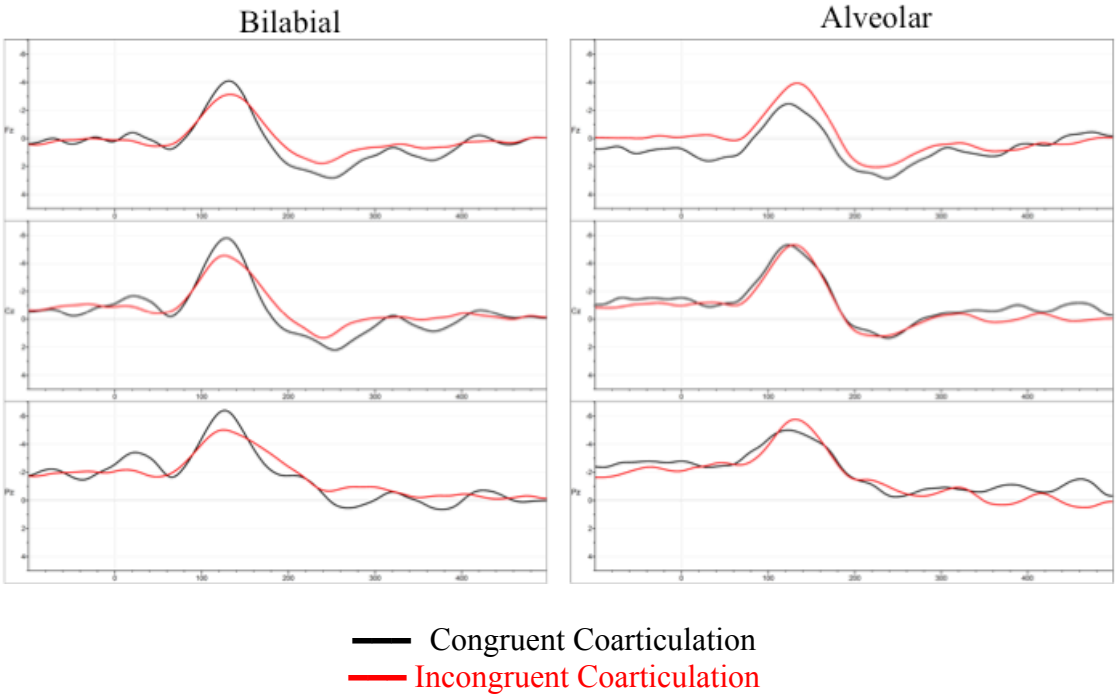
Oral:Nasal - An ANOVA revealed that mean amplitude in the PMN response was significantly larger to oral onsets (M = -4.49 µV, SD = .21) than to nasal onsets (M = -3.77 µV, SD = .10), $F(1,55) = 18.62$, $p > .001$.

Figure 18 Grand average waveforms for all conditions organized by onset type: oral vs. nasal stops, at sites Fz, Cz, and Pz.



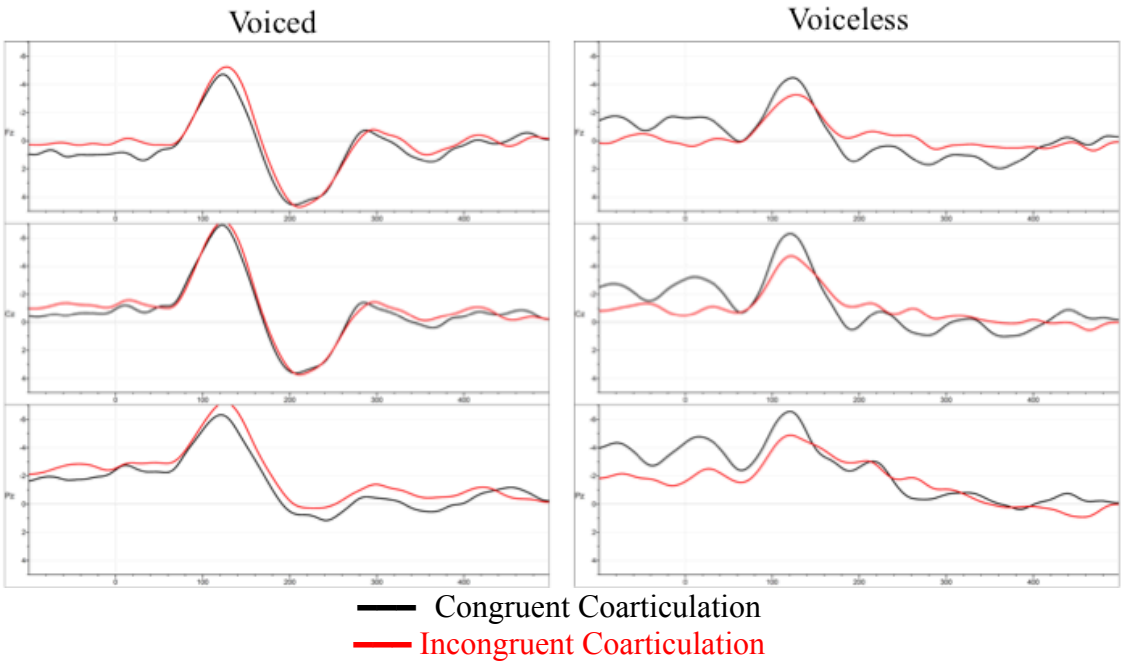
Bilabial:Alveolar - Results of an ANOVA test revealed that there was no significant difference in PMN amplitude when onset data was organized by place of articulation, $F(1,55) = 18.48$, $p > .05$.

Figure 19 Grand average waveforms for all conditions, grouped by place of articulation of the onset, shown for sites Fz, Cz, and Pz.



Voiced:Voiceless - The final ANOVA performed across onset type grouping was done on all onsets depending upon whether they were voiced or voiceless. Nasals were excluded from the analysis due to their acoustic and temporal differences with respect to oral plosives. Results revealed a significant main effect of voicing on the PMN amplitude, $F(1,55) = 33.85$, $p < .001$. A *t*-test revealed that the mean amplitude to voiceless onsets was significantly larger (M = -2.37 µV, SD = 0.16) than voiced onsets (M = -1.99 µV, SD = 0.23), $t(1) = -5.18$, $p < .001$.

Figure 20 Grand average waveforms for all conditions, organized such that onsets were either voiced or voiceless. Nasal onsets were not included in averages. Waveforms shown for sites Fz, Cz, and Pz.



5.0 Discussion

The present experiment investigated the effects of congruent and incongruent coarticulation on event-related potentials as reflected through event-related brain potentials. Neural responses were recorded to auditory stimuli containing congruent or incongruent coarticulatory information in the transition from onset to nucleus of systematically varying monosyllabic English words. Stimuli varied across several conditions, including onset consonant identity (oral/nasal, bilabial/alveolar, voiced/voiceless) and vowel type (height and backness). As reported in Section 4.0, main effects and interactions were observed regarding three neural signatures: the N100, the P200, and the PMN. These results are consistent with and elaborate upon those reported

in Archibald & Joanisse (2011) and Arbour (2012). The results from this study delve deeper into the specific effects that degree of difference between two articulations has on the PMN. The results presented here reflect the observation that physical properties of sounds produce different degrees of violation in a way that is directly related to the degree of difference in acoustic space between two phonemes. The following is an elaboration of the comparisons discussed above and a discussion of their implications.

5.1 N100 and P200

The N100 response is a neural response elicited by auditory stimuli, and is interpreted as being a precursor to speech processing (Steinhauer & Connolly, 2008). The amplitude of the response is reliably larger to syllables in word-onset position than in word-medial position (Steinhauer & Connolly, 2008). Thus, the results reported above focus on the N100 response amplitude as predicted by onset type. Results demonstrated that the N100 is sensitive to acoustic differences between phonemes. N100 responses were significantly different between oral and nasal stops in the onset position. This is likely due to the spectral and temporal differences between the phonemes used in this study (Arbour, 2012). Nasal stops are sustained and periodic, whereas oral stops are aperiodic, temporally brief, and emit a burst of noise (Stevens, 2002). Additionally, the N100 amplitude to voiced plosives was smaller than to voiceless plosives. This again can be attributed to phonetic differences, such as the earlier voice onset time of voiced stops as compared to voiceless stops (Reetz & Jongman, 2011). Finally, there was no significant difference observed between the N100 amplitude and place of articulation.

This could be due to the fact that stops that differ in place of articulation are the same in terms of spectral and temporal properties.

The P200 has previously been tied to the N100 response in speech processing studies, where both are viewed as part of a larger complex (Steinhauer & Connolly, 2008). The P200 behaved similarly to the N100 in regard to voicing. However, opposing effects were found for place of articulation and oral versus nasal onsets. These results could be attributed to difference in voice onset time between voiced and voiceless stops. There was also a significantly smaller P200 amplitude to incongruent conditions than to congruent conditions. This could be a regressive effect of the PMN on preceding components. Additionally, as discussed in Arbour (2012), this difference could reflect the fact that coarticulatory cues are detected early in stop consonants (Feng et al., 2011).

5.2 PMN

Early work on the PMN identified its sensitivity to between-category phonological mismatches (Connolly & Phillips, 1994; Newman & Connolly, 2009). Archibald & Joanisse (2011) were the first to demonstrate the PMN's sensitivity to within-category subphonemic mismatches, with these results being confirmed by Arbour (2012). The aim of the present study was to investigate how specific acoustic differences in incongruous coarticulation would affect the PMN response. Results revealed the sensitivity of the PMN and its relationship to the degree of phonological difference between two segments.

5.2.1 Coarticulatory Congruity

Results confirmed reports (Archibald & Joanisse, 2011; Arbour, 2012) that the PMN was larger to incongruent coarticulatory cues as opposed to target words carrying

congruent coarticulation. The demonstrated amplitude difference again reinforces the notion that the PMN is not only sensitive to within-category phonological violations, but also to *between*-category violations. These results lend support to the idea put forth by McQueen & Cutler (1997) that coarticulatory variation is "lawful", and that listeners have some implicit knowledge of these patterns.

The key factor that discriminates the present experiment from its predecessors (Archibald & Joanisse, 2011; Arbour, 2012) is the particularity with which comparisons were made between types of subphonemic incongruities. The following sections discuss the effects of consonant onsets and vowel distance as they modulated the PMN response.

5.2.2 Onset Type

While Arbour (2012) compared and found differences in the PMN to fricative versus stop onsets, the present study investigated effects within the stop category. Recall that onsets in the present study were drawn from the English anterior stop series /p, b, m, t, d, n/. Onsets were organized into natural phonological classes and analyzed accordingly. The phonemes were classed together by place of articulation, voicing, and whether they were orally or nasally produced. The only grouping that did not yield significant results was based upon the place of articulation of the onset.

A significant effect of voicing was found among the oral stops in the series. The PMN amplitude was significantly larger when incongruous coarticulation followed a voiceless stop than a voiced stop. This can be explained and attributed to the aspiration of voiceless stops in English. Voiceless oral stops in Canadian English are reliably aspirated. Thus, the stop burst is followed by a release of fricative noise that closely

resembles the glottal fricative /h/. Wright (2004) states that coarticulatory cues are more readily available in fricative noise than from a stop burst. The aspiration that follows voiceless stops acts as this fricative noise, carrying a greater amount of coarticulatory information. This result supports the proposition (Archibald & Joanisse, 2011) that /h/ would carry the greatest amount of coarticulatory information, due to the open configuration of the vocal tract during its production.

The PMN was also differentially modulated based upon whether the word onset was orally or nasally produced. The PMN response was significantly larger when incongruent coarticulation followed oral consonants rather than nasal consonants. Although nasals are more sustained and periodic, they do not allow for shaping of the vocal tract in the way that oral stops do. Perhaps the vocal tract is modified according to the following vowel during production of a nasal stop, but no air is passing through the oral tract, so oral configuration cues are not being realized audibly.

5.2.3 Vowel Type

The present study aimed to carefully balance and compare the phonological shape of the stimuli. One of four corner vowels of English /i, u, æ, ɑ/ were used as the nucleus of each stimulus word. Selecting this set of vowels allowed for maximal distance between vowels that varied along both dimensions of height and backness (Liljencrants & Lindblom, 1972). Consequently, moving from the vocal tract configuration associated with one vowel as compared to another corner vowel would necessitate the greatest amount of modulation of the articulators. Arbour (2012) found no difference in coarticulatory cues in words containing high vowels. The present experiment aimed to

perform a more complex comparison by examining both high and low vowels in the same experiment.

Vowel identity was found to modulate the PMN in incongruent conditions. Results showed that the amplitude of the PMN response was modulated as a factor of the degree of difference between the post-splice vowel and the vowel whose identity was cued in the onset phoneme. For example, if the coarticulated vowel and post-splice vowel differed in only one dimension (height or backness), the amplitude of the PMN response was attenuated compared to if the two vowels differed along two dimensions (i.e., height *and* backness).

Looking specifically at the condition in which onset cues identified the vowel /æ/ provides an example of the points being made. The vowel /æ/ differs along only one dimension compared to the vowel /i/ (height) and the vowel /ɑ/ (backness). The vowel /æ/ differs from the vowel /u/ in more than one dimension: height and backness (and rounding). The waves for incongruities involving /i/ and /ɑ/ are clustered closer together than the wave for the incongruous condition involving /u/. These results demonstrate the effects that acoustic properties have on listeners' perception of coarticulation in speech.

5.3 Implications

The results reported here demonstrate the direct relationship between two quantitative aspects of phonology. The PMN is directly modulated by the degree of difference between the acoustic properties of phonemes. This finding can be compared to a neurophysiological cortical recording study from Bouchard, Mesgarani, Johnson & Chang (2013). This study had participants produce several CV syllables while electrodes

placed on the surface of the cortex recorded electrical activity. These data were used in an attempt to locate and map phonemes onto their related areas of the sensorimotor cortex. Results of Bouchard et al. (2013) succeeded in locating an area on the sensorimotor cortex that corresponded to each articulator that was activated during recording.

These results along with the results of the present experiment support a notion of speech perception that is particularly sensitive to the articulatory properties of speech sounds. Results such as these fit into a model of speech perception that is sensitive to fine acoustic detail as it occurs in real time, and one that can integrate higher-level information to achieve ultimate word recognition (i.e., a top-down model). The continuous acoustic integration hypothesis as well as the Merge model of spoken word recognition provide the best account of the current results. Participants' neural responses demonstrated their sensitivity to subphonemic cues, but their behavioural results indicated that word recognition was ultimately not affected. This suggests that speech perception is a fluid process that is resilient to degraded input. Listeners' brains can recover from "imperfections" in the speech stream in such a way that it does not preclude further processing.

6.0 Conclusion

EEG is a valuable tool that has the temporal resolution necessary to detect such fine-grained acoustic details in real-time on the millisecond scale. The balanced set of stimuli allowed for careful comparison between and within sound classes that was not possible in previous experiments. The current experiment builds upon the interpretation

of the PMN response to within-category phonetic violations. It demonstrates listeners' sensitivity to the measurable acoustic differences between sound classes and within phonemes. The findings presented here are consistent with those in Archibald & Joanisse (2011) and Arbour (2012). The present study also confirms notions about speech processing that focus on continuous integration of fine phonetic detail, such as those proposed by Gow (2007). Subphonemic cues can indeed affect spoken word processing, and facilitate word recognition.

6.1 Future Directions

Marslen-Wilson (1993) states that oftentimes, monosyllabic words differ from at least one other monosyllabic word by just one feature. Recall, the present study focused on subphonemic features using monosyllabic words, finding that even the smallest featural differences are meaningfully processed. Perhaps the inherent monosyllabicity of the experimental stimuli is what necessitated listeners to meaningfully attend to subphonemic cues. The present results would be strengthened by further work that focused on incongruent coarticulation in word-medial position in polysyllabic words. It is questionable whether these conditions would elicit a PMN, if the words were lengthy enough, eliminating lexical candidates to the extent that there are only a few word-completion possibilities. Perhaps in this setting listeners would not need to rely so heavily on subphonemic cues while processing.

References

Allopenna, P. D., Magnuson, J. S., & Tananhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Anderson, H. (1973). Abductive and deductive change. *Language, 49*, 765-793.

Arbour, Jessica, "The Dynamic Role of Subphonemic Cues in Speech Perception: Investigating Coarticulatory Processing Across Sound Classes" (2012). *Open Access Dissertations and Theses*. Paper 7503. Available from http://digitalcommons.mcmaster.ca/opendissertations/7503

Archibald, L. M., & Joanisse, M. F. (2011). Electrophysiological responses to coarticulatory and word level miscues. *Journal of Experimental Psychology: Human Perception and Performance, 37*(4), 1275-1291.

Boersma, P., & Weenink, D. (1995). Praat (Version 5.3.56) [Software]. Available from www.praat.org

Brain Vision Analyzer (Version 2.0.4) [Software]. Brain Products, GbmH.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Cole, R. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 1,* 153-156.

Cole, R., & Jakimik, J. (1978). Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of information processing* (pp. 67-116). New York: Academic Press.

Connolly, J. F., & Phillips, N. A. (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *Journal of Cognitive Neuroscience, 6*(3), 256-266.

Connolly, J. F., Byrne, J. M., & Dywan, C. A. (1995). Assessing adult receptive vocabulary with event-related potentials: An investigation of cross-modal and cross-form priming. *Journal of Clinical and Experimental Neuropsuchology, 17*(4), 548-565.

Connolly, J. F., Service, E., D'Arcy, R. C. N., Kujala, A., & Kimmo, A. (2001). Phonological aspects of word recognition as revealed by high-resolution spatio-temporal brain mapping. *Cognitive Neuroscience and Neuropsychology, 12*(2), 237-243.

Davies, Mark. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at http://corpus2.byu.edu/glowbe/.

Elman, J. L., & McClelland, J. L. (1988). Cognitive Penetration of the Mechanisms of Perception: Compensation for Coarticulation of Lexically Restored Phonemes. *Journal of Memory and Language, 27,* 143-165.

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences*, 371-404.

Feng, Y., Hao, G. J., Xue, S. A., & Max, L. (2011). Detecting anticipatory effects in speech articulation by means of spectral coefficient analyses. *Speech Communications, 53*, 842-854.

Fleming, E. (1997). Phonetic detail in phonology: Towards a unified accout of assimilation and coarticulation. In K Suzuki & D. Elzinga (Eds.), *Proceedings of the 1995 Southwestern Workshop I Optimality Theory (SWOT)*. Tucson: University of Arizona.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics, 28*, 267-283.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language, 45*, 133-159.

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 163-179.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65*(4), 575-590.

Gow, D. W., & McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation. *Papers in laboratory phonology, 9*, 173-200.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203-205.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potiential (ERP). *Annual review of psychology*, *62*, 621-647.

Lahiri, A., & Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38,* 245-294.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*(4), 839-862.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

McQueen, J. M., & Cutler, A. (1997). Cognitive processes in speech perception. In W. J. Hardcastle & J. Laver. (Eds.). *The handbook of phonetic sciences* pp. 566-585. Blackwell Reference Online.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance, 25*(5), 1363-1389.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71.

Marslen-Wilson, W. D. (1993). Issues of process and representation. In G. T. M. Altmann, & R. Shillcock (Eds.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 187-210). Cambridge: MIT Press.

Martin, B. A., Tremblay, K. L., & Stapells, D. R. (2007). Principles and applications of cortical auditory evoked potentials. In J. Eggermont, M. Don & R. Burkard (Eds.), *Auditory evoked potentials: Basic principles and clinical applications* (pp. 482-507). Philadelphia: Lippincott, Williams & Wilkins.

Newman, R.L., & Connolly, J.F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom-up and top-down effects on spoken word processing. *Biological Psychology, 80*, 114-121.

Newman, R. L., Connolly, J. F., Service, E., & McIvor, K. (2003). Influence of phonological expectation during a phoneme deletion task: Evidence from event-related brain potentials. *Psychophysiology, 40*, 640-647.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299-370.

R Core Team. (2012). R: A language and environment for statistical computing (Version 0.98.994) [Software]. Available online from http://www.R-project.org/

Raphael, L. J. (2008). Acoustic cues to the perception of segmental phonemes. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (pp. 182-206). Oxford: Blackwell Publishing.

Recasens, D., Pallarès, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America, 102*(1), 544-561.

Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, production, acoustics, and perception* (Vol. 34). John Wiley & Sons.

Smits, R. (2001). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception & Performance, 27,* 1145-1162

Steinhauer, K., & Connolly, J. F. (2008). Event-related potentials in the study of language. In B. Stemmer & H. Whitaker (Eds.), *Handbook of the cognitive neuroscience of language* (pp. 91-104). New York: Elsevier.

Stevens, K. N. (2002). Toward a model of lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*(4), 1872-1891.

Whalen, D. H. (1991). Perception of the English /s/ - /ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *Journal of the Acoustical Society of America, 90*(4), 1776-1785.

Wright, R. (2004). A review of perceptual cues and cue robustness In B. Hayes, R. Kirchner, & D. Steriade (Eds.). *Phonetically based phonology* pp. 34-57. Cambridge University Press.

Yeni-Komshian, G., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 70*(4), 966-975.

Appendix

A. Full list of stimuli organized into sets by onset and coda.

beep	
	bop

beat	boot
bat	bought

beak	
back	

	boob
	bob

bead	
bad	

bag	bog

peep	poop
pap	pop

pat	pot

peak	
pack	

pad	pod

deep	

	dot

deke	duke
	dock

dab	

deed	dude
dad	

	dog

tap	top

	toot
tat	tot

teak	toque
tack	talk

	tube
tab	

teed	
	Todd

tag	

map	mop

meet	moot
mat	

meek	
Mac	mock

	mob

mead	mood
mad	

mag	

nap	

neat	newt
gnat	not

knack	knock

nab	knob

need	nude
	nod

nag	nog

B. List of conditions with trial counts.

	Pre-splice Vowel	Post-splice Vowel	Trial count (per participant)
Congruent	i	i	38
	u	u	26
	æ	æ	54
	ɑ	ɑ	66
Incongruent	i	u	40
	i	æ	24
	i	ɑ	24
	u	i	32
	u	æ	36
	u	ɑ	18
	æ	i	38
	æ	u	34
	æ	ɑ	48
	ɑ	i	40
	ɑ	u	22
	ɑ	æ	40
Unrelated	[congruent]	[congruent]	75

C. Letter of informed consent

McMaster University
Inspiring Innovation and Discovery

LETTER OF INFORMATION / CONSENT

A study of speech processing in the brain:
The effect of coarticulation on spoken word recognition

Principal Investigator: Samantha Kramer
Department of Linguistics and Languages
McMaster University
Hamilton, Ontario, Canada
E-mail: kramerse@mcmaster.ca

Faculty Supervisors: Dr. John F. Connolly
Department of Linguistics and Languages
McMaster University
Hamilton, Ontario, Canada
(905) 525-9140 ext. 27095
E-mail: jconnol@mcmaster.ca

Dr. Anna L. Moro
Department of Linguistics and Languages
McMaster University
Hamilton, Ontario, Canada
(905) 525-9140 ext. 23762
E-mail: moroal@mcmaster.ca

Purpose of the Study

You are invited to take part in this study on the perception of speech, specifically word recognition, in the brain. I, as a Master's student in the Cognitive Science of Language program, want to observe word recognition processes by recording event-related brain potentials from native English speakers. With this project, I am hoping to gain insight into the processes underlying spoken word recognition, and contribute this knowledge to an existing model of spoken word recognition.

Procedures involved in the Research

You will be seated in front of a computer screen for the duration of this experiment. In order to measure your neural activity, you will be fitting with an electrode cap.

Electroencephalography (EEG) is a non-invasive neuroimaging technique used to measure event-related brain potentials (ERPs). Recording EEG requires you to wear the cap that holds electrodes on the surface of your scalp. During experimental trials, you will be presented visually with a word. Once the word has been presented, you will hear an auditory word through headphones. Once the word has been played, you will be asked to indicate whether or not the word you heard matched the word you read. The entire study will be conducted in one session lasting approximately 2 hours, and you will be given 20 short breaks at regular intervals. You will compensated with course credits through the SONA system. These credits will be applied to your account after you complete the experiment.

Potential Harms, Risks or Discomforts:

It is not likely that there will be many harms associated with this study. However, electrolyte gel, which improves the conductivity of the electrodes, will be applied to your scalp. After the session is complete, you will be given the opportunity to wash the gel from your hair. Additionally, you may become fatigued during the session.

You do not need to answer questions that you do not want to answer or that make you feel uncomfortable. You can withdraw (stop taking part) at any time. Described below are the steps I am taking to protect your privacy.

Potential Benefits

The research will not benefit you directly. I hope to learn more about how the brain processes subtle cues in speech. I hope that what is learned as a result of this study will help to better the understanding of how speech sounds are processed in real-time. Long-term, these results could aid in the advancement of automatic speech recognition technology.

Confidentiality

You are participating in this study confidentially. I will not use your name or any information that would allow you to be identified. No one but me will know whether you participated unless you choose to tell them.

The recorded data will be kept on a computer that will be protected by a password. Once the study has been completed, the data will be destroyed after one year.

Participation and Withdrawal

Your participation in this study is voluntary. It is your choice to be part of the study or not. If you decide to be part of the study, you can decide to withdraw, at any time, even after signing the consent form or part-way through the study. If you decide to withdraw,

there will be no consequence to you. In cases of withdrawal, any data you have provided will be destroyed unless you indicate otherwise.

Information about the Study Results

I expect to have this study completed by the end of May 2014. If you would like a brief summary of the results, please let me know how you would like it sent to you.

Questions about the Study

If you have questions or require more information about the study itself, please contact me.

This study has been reviewed by the McMaster University Research Ethics Board and received ethics clearance.
If you have concerns or questions about your rights as a participant or about the way the study is conducted, please contact:

McMaster Research Ethics Secretariat
Telephone: (905) 525-9140 ext. 23142
c/o Office of Research Services
E-mail: ethicsoffice@mcmaster.ca

CONSENT

I have read and understood the information letter about the study being conducted by Samantha Kramer of McMaster University, and have been given a copy of the information letter. I have had the opportunity to ask questions about my involvement in this study, and to receive any additional details and clarification. I understand that I may withdraw from the study at any time.

Name of Participant

Signature

Date

D. Participant debriefing form

PARTICIPANT DEBRIEFING FORM

A study of speech processing in the brain:
The effect of coarticulation on spoken word recognition

Thank you for your participation in this study!

The general purpose of this research is to examine neural responses to spoken words. Specifically, I am examining a phenomenon called *coarticulation*. Coarticulation is a subphonemic process that arises during the transition between articulatory gestures during speech. For example, there are coarticulatory differences in the /k/ sound in the words “keep” and “coop”, even though both these words start with a /k/ sound. These differences are due to anticipating the following vowel. Try preparing to say “keep” and “coop” yourself. You’ll notice that when you’re about to say the word “keep”, your mouth is spread wide. When you’re about to say “coop”, your mouth is rounded.

There are claims that coarticulatory cues aid speakers in word recognition and that using these cues can speed up processing. This study is interested in what happens when these cues are misleading. A neural marker called the *phonological mapping negativity* (PMN) has been observed when listeners hear a coarticulatory mismatch. I am hoping that the conditions in this study will elicit the same response. If a PMN is observed, this provides support for the fact that coarticulation is not disregarded by listeners as random noise, but that it is processed systematically by listeners.

We invited participants who were typically developed adult native speakers of English with no history of medical or neurological conditions. In this study, you were asked to take part in a speech perception task, which involved reading words on a computer screen, then hearing spoken words and deciding whether what you heard matched what you read. I am interested in the tokens where the words matched but the spoken word contained a coarticulatory mismatch. The results from this study will provide insight into how humans process subphonemic cues such as coarticulation. We hope that empirical data from this experiment will contribute to unanswered questions about the perception of variation in speech and how it is processed in the brain.

You may obtain additional information about the results of the study by contacting the principle investigator (Samantha Kramer) at kramerse@mcmaster.ca. If you have any additional concerns or questions, you may contact the McMaster Research Ethics Board Secretariat at (905) 525-9140 Ext. 23142.

E. Screening form to ensure eligibility of all participants

SCREENING FORM
Participant code: _____ Date of birth: _____ Test date: _____
Handedness: □ Right □ Left □ Ambidextrous Sex: □ Male □ Female
Highest level of education: _____
Languages in order of fluency: 1._____
2._____ 3. _____
4. _____
If English is not your first language: How old were you when you learned English?
_____
If you were not born in Canada: How old were you when you moved to Canada?
_____
History of substance abuse:
_____
Is your hearing and vision normal? □ Yes □ No
If not, please describe:
_____
Have you ever had any perceptual (colour blindness) learning or language problems?
□ Yes □ No
If yes, please describe (age, length, recovery):
_____
Have you ever had any neurological, psychological or psychiatric problems?
□ Yes □ No
If yes, please describe (age, length, recovery):
_____
Have you ever had a head injury, seizures, coordination problems or major surgeries?
□ Yes □ No
If yes, please describe (age, length, recovery):
_____
Have you ever lost consciousness, had any fainting spells, paralysis or dizziness?
□ Yes □ No
If yes, when and for how long?
_____
Are you presently taking any medication? □ Yes □ No
If yes, which one(s)?
_____
_____
Have you recently taken any medication? □ Yes □ No
If yes, which one(s), and when?
_____

Do you consume the following?

		How often?
Alcohol	□Yes □No	
Cigarettes	□Yes □No	
Drugs	□Yes □No	

Edinburgh Handedness Inventory[1]

Please indicate with a check (✓) your preference in using your left or right hand in the following tasks.
Where the preference is so strong you would never use the other hand, unless absolutely forced to, put two checks (✓✓).
If you are indifferent, put one check in each column (✓| ✓).
Some of the activities require both hands. In these cases, the part of the task or object for which hand preference is wanted is indicated in parentheses.

Task / Object	Left Hand	Right Hand
1. Writing		
2. Drawing		
3. Throwing		
4. Scissors		
5. Toothbrush		
6. Knife (without fork)		
7. Spoon		
8. Broom (upper hand)		
9. Striking a Match (match)		
10. Opening a Box (lid)		
Total checks:	LH =	RH =
Cumulative Total	CT = LH + RH =	
Difference	D = RH – LH =	
Result	R = (D / CT) × 100 =	
Interpretation: (Left Handed: R < -40) (Ambidextrous: -40 ≤ R ≤ +40) (Right Handed: R > +40)		

[1] Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsycholol gia, 9, 97-113.