View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Louisville

University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2017

Neural coding of natural and synthetic speech.

Allison Brown
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

Part of the Communication Sciences and Disorders Commons

Recommended Citation

Brown, Allison, "Neural coding of natural and synthetic speech." (2017). *Electronic Theses and Dissertations.* Paper 2648.
https://doi.org/10.18297/etd/2648

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

NEURAL CODING OF NATURAL AND SYNTHETIC SPEECH

By

Allison Brown

B.S.- University of Kentucky, Lexington Kentucky, 2015

A Thesis
Submitted to the Faculty of the
School of Medicine of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science
in Communicative Disorders

Department of Otolaryngology Head and Neck Surgery and Communicative Disorders
University of Louisville
Louisville, Kentucky

May 2017

© 2017
Allison Brown
All rights reserved

NEURAL CODING OF NATURAL AND SYNTHETIC SPEECH

By

Allison Brown

B.S.- University of Kentucky, Lexington KY, 2015

A Thesis Approved on

April 21, 2017

by the following Thesis Committee:

____

Teresa Pitts, Ph.D., Thesis Director

____

Sharon E. Miller, Ph.D.

____

Alan Smith, Ph.D.

.

DEDICATION

To my thesis committee and to my family, thank you.

ACKNOWLEDGMENTS

I first and foremost want to thank my thesis advisor, Dr. Sharon Miller. Who willingly accepted me as an inexperienced student and not only helped me produce this thesis, but also developed me into a researcher, a writer, and a lifelong learner. The patience and guidance she has provided me are tools I will carry with me for the entirety of my professional career. None of this would be possible if not for you.

On the first day of orientation, I expressed to Dr. Alan Smith that it was my personal goal to complete a thesis during my time at UofL. Thank you, Dr. Smith, for holding me accountable to this goal and for providing me with all the tools I needed to accomplish it.

It is because of Dr. Teresa Pitts that I believe anything is achievable once you set your mind to it. Dr. Pitts, you are a role model to me within our field and in life. Thank you for believing in me.

I cannot go without thanking my classmates, it is to them that I owe my sanity, and my hope in the successful future of our profession. I am fortunate to have learned alongside you for two years.

Most importantly, to my family. I am eternally grateful to them for enabling me in all that I do. Thank you for listening to me, investing in me, and standing by me. I can never thank you enough.

ABSTRACT

NEURAL CODING OF NATURAL AND SYNTHETIC SPEECH

Allison Brown

April 21, 2017

The present study examined whether natural and synthetic speech are differentially encoded in the auditory cortex. Auditory event-related potential (ERP) waveforms were elicited by natural and synthetic fricative-vowel stimuli (/sɑ/ and /ʃɑ/) in a passive listening paradigm in adult listeners with normal hearing. ERP response components were compared across conditions. The results indicated that peak latencies to natural speech were significantly earlier than those to synthetic speech. Natural speech also produced significant electrode hemisphere site effects, whereas synthetic speech activated left, midline, and right electrode hemisphere sites equally.   Overall, the results suggest that cortical processing of natural and synthetic speech activates distinct neural systems which has important clinical implications for the speech-language pathology field.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

Overview

Speech perception involves the mapping of an acoustic signal from a speaker to mental representations of phonemes, words, and sentences in a listener. This thesis project examines the neural mechanisms underlying speech perception and, specifically, investigates whether the neural coding of fricative speech sounds is affected by whether they are naturally produced by a human talker or synthesized by a computer. This project used electroencephalography (EEG) measures to compare how stimulus characteristics affect cortical responses in listeners with normal hearing.

Background

Speech Perception

Accurate speech perception is the foundation for successful human communication. The process by which the brain derives meaning from a dynamic, acoustically variable speech signal is of immense interest to many, including speech language pathologists and audiologists. Proper perception of naturally produced spoken language requires a listener to perceptually map the incoming, variable acoustic speech signal onto phonetic categories, access words stored in the mental lexicon, and combine words in a semantically meaningful way to compute the correct meaning of an utterance (M. S. Gazzaniga, 2009). Multiple theories of how a listener derives meaning from the auditory speech signal exist. The theories can be broadly classified into bottom-up, top-down, or interactive models (Figure 1).

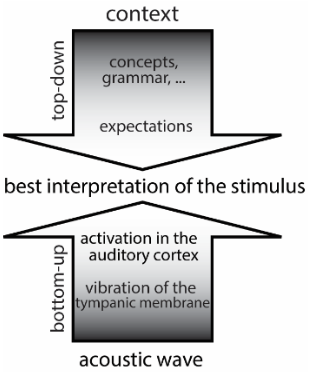


Figure 1. Schematic representation of the bottom-up and top-down components involved in auditory speech perception.

Bottom-up theories, or abstract approaches, of speech perception posit that speech perception proceeds in a serial fashion whereby listeners first need to perceive individual phonemes in order for lexical access to occur (Oden & Massaro, 1978). In contrast, top-down theories of speech perception suggest that context and lexical knowledge influence phonetic perception (see Pisoni and Levi, 2007 for a review) . Finally, hybrid approaches suggest that speech perception is an interaction of bottom-up and top-down effects with both feedforward and feedback mechanisms (e.g. McLelland and Elman, 1986).

Natural versus Synthetic Speech Perception

How easily and accurately a listener accesses words stored in the mental lexicon is known to be affected by properties of the speech signal (Pisoni & Levi, 2007), and whether the speech is naturally produced by a human or synthetically produced by a computer (Luce, Feustel, & Pisoni, 1983). Naturally produced speech contains numerous suprasegmental, or prosodic, features, such as duration, intonation, and stress, that are superimposed on phonemes, words, and sentences; these prosodic features play a key role in helping the listener parse running speech and contribute to accurate speech understanding (see Cutler, Dahan, & van Donselaar, 1997 for a review). In contrast, synthetic speech lacks the prosodic features of natural speech, and while it tries to mimic the frequency, amplitude, and source characteristics of natural speech, it does not contain any of the same inherent variability or acoustic-phonetic cue redundancies (Borden Gloria J, 2011; Greene, 2005).

Previous behavioral studies have examined whether differences in synthetic and naturally produced speech affect different aspects of speech perception including segmental intelligibility, lexical decision making, word recall, and sentence comprehension (Clark, 1983; Luce et al., 1983; Nusbaum, Dedina, & Pisoni, 1984; Pisoni, 1981). Early work by Clark (1983) examined segmental intelligibility of naturally produced and synthetic consonant-vowel-consonant (CVC) stimuli and consonant-vowel (CV) stimuli in the presence of white noise. The results indicated that perception of synthetic consonants was significantly affected by background noise relative to the perception of naturally produced speech, and the effect was most pronounced for fricative and stop consonants. This finding was supported by Nusbaum, Dedina and Pisoni (1984) who investigated whether the lack of acoustic-phonetic redundancy accounted for the poorer segmental intelligibility of synthetic speech in background noise. Nusbaum et al. (1984) measured intelligibility of synthetic and naturally produced CV stimuli in background noise and examined the consonant confusion matrices for each. The findings suggested that naturally produced speech was perceived more accurately than synthetic speech, but, for some consonants, the pattern of errors differed substantially for the synthetic and naturally produced stimuli. For other consonants, the pattern of errors was similar for the synthetic and naturally produced speech. The authors theorized that when there was a difference in error patterns across stimuli, the minimal cues available for the synthetic speech in noise were actually misleading and incorrect, but when the patterns of errors were similar, the noise reduced the redundancy similarly across stimuli. Thus, they concluded that the acoustic cue structure of synthetic speech can be misleading and lacks the same acoustic-phonetic cue redundancy of natural speech.

In addition to differences in segmental intelligibility, Pisoni (1981) investigated whether processing synthetic speech requires more cognitive resources than natural speech by using a speeded lexical decision task. In the task, listeners were presented with naturally-produced or synthetic word and non-word stimuli and had to determine if the stimulus was a real word or not. The results indicated that listeners' reaction times were significantly longer for the synthetic speech stimuli than for the naturally produced stimuli, regardless if a word or non-word stimulus was presented. Pisoni (1981) concluded that the longer reaction times for the synthetic speech likely indicated that listeners were using more cognitive resources to process the acoustic-phonetic structure prior to any higher order processing. The results could not be accounted for by listeners being more familiar with natural speech as the effect was consistent, even with repeated exposure to the synthetic stimuli (Slowiaczek & Pisoni, 1982).

Luce, Feustel and Pisoni (1983) further investigated whether the increased cognitive processing demands for synthetic speech constrained short-term memory processing and subsequent transfer to long term memory. Listeners were asked to recall lists of naturally produced and synthetic words with and without a digit pre-loading task. For the digit pre-loading task, listeners had to memorize 0 to 6 numbers and recall them in order before completing the word recall task. Without digit pre-loading, subjects' recall for naturally produced words was more accurate than for synthetically produced words. In addition, subjects had significantly more errors for the synthetic stimuli where they recalled words not present on the stimulus lists. The same trend was observed for the digit pre-loading condition, but with greater number of errors overall. The authors hypothesized that the results indicated that synthetic word lists were harder to maintain, process, and store than naturally produced words.

Behavioral evidence suggests that synthetic speech is difficult to understand and requires greater cognitive capacity to process (Luce et al., 1983; Nusbaum et al., 1984; Pisoni, 1981). However, synthetic speech is easy to produce and is widely used in today's technology. Synthetic speech is also prevalent in the speech-language pathology domain. For example, augmentative and alternative communication (AAC) devices, devices that enable persons with speech production impairments to communicate, primarily use synthetic speech for verbal communication. Thus, it is important to better understand why differences in behavioral comprehension and intelligibility exist between natural and synthetic speech.

Fricative Perception

Behavioral studies of naturally produced versus synthetic speech indicate that perception is less efficient for synthetic speech, and that synthetic fricative speech sounds are often the most subject to errors (Clark, 1983; Luce et al., 1983; Nusbaum et al., 1984).

The acoustic characteristics of fricatives could account for why they are often subject to misperception. The speech signal consists of consonant and vowel sounds whose production can be described using a source-filter model (Fant, 1960). For vowels, the vibrating vocal folds are the source, producing a complex periodic wave. For consonants, the source is either aperiodic noise or a both aperiodic noise and the harmonic spectrum from the vibrating vocal folds (Johnson, 2003). In the source-filter theory model, the vocal tract acts as an acoustic filter, shaping the acoustic output from the source (Fant, 1960). Unlike vowels that are produced with a relatively open vocal tract, consonants are produced with a constriction in the vocal tract. Where this constriction occurs is referred to as the place of articulation, and how the constriction occurs is referred to as the manner of articulation (Johnson, 2003). Fricative speech sounds are produced when turbulent noise is produced and escapes past a narrow constriction in the vocal tract (Johnson, 2003). In general, relative intensity is lower for fricatives than vowels, and fricatives lack the same well defined formant structure as vowels (Borden, Harris, & Raphael, 2003). Fricatives are typically classified as sibilant (/s/, /z/, /ʒ/, /ʃ/) or non-sibilant (/f/, /v/, /ɵ/, /ð/). The English voiceless sibilant contrast /s/-/ʃ/ will be the focus of this thesis. The /s/-/ʃ/ contrast differs in place of articulation, with /s/ classified as an alveolar and /ʃ/ a palato-alveolar. The contrast differs in peak spectral energy, with /s/ usually having a spectral peak near 4 to 8 kHz and /ʃ/ having spectral peak energy around 2 to 5 kHz (Ladefoged, 1962; Stevens, 1998).

In listeners with normal hearing, perception of fricatives is known to depend on access to the dynamic transition cue and the spectral shape of the frication noise (Zeng & Turner, 1990). It is possible synthetic fricatives are subject to more misperceptions because the acoustic cues for fricatives are less robust compared to other speech sounds, and these already weak cues may become more easily distorted during speech synthesis.

Behavioral versus Neurophysiological Approach

Behavioral studies of natural versus synthetic speech perception suggest that the two types of stimuli are not processed similarly when using reaction time and percent correct measures (Luce et al., 1983; Nusbaum et al., 1984; Pisoni, 1981), and synthetic fricative speech sounds were found to be subject to more misperceptions than other types of consonant sounds (Clark, 1983). Information-processing theory (Atkinson & Shiffrin, 1971) posits that the accuracy and timing of behavioral responses is related to the difficulty and ease of processing, suggesting that synthetic speech uses more and/or different cognitive resources to process. While behavioral studies can inform us that differences in performance exist, they cannot define what the underlying neural processes are that support the observed differences. Neurophysiological and neuroimaging measures can examine the cortical mechanisms underlying the processing differences.

Neuroimaging methods have emerged as powerful tools for investigating the neural mechanisms underlying speech perception. Electroencephalography (EEG), Magnetoencephalography (MEG), and Functional Magnetic Resonance Imaging (fMRI) measures are now commonly used to examine speech and language processing in the cortex (M. Gazzaniga & Mangun, 2014). The different methods have strengths and weakness when it comes to studying language due to trade-offs in spatial and temporal resolution across techniques. fMRI measures the hemodynamic blood flow differences across tasks and has exquisite spatial resolution. However, the blood flow response is quite sluggish and on the order of seconds, so it does not have the temporal precision to respond to the dynamic changes in the speech signal that occur at a much faster rate. EEG and MEG, on the other hand, have exquisite temporal resolution, but poorer spatial resolution than fMRI. Because EEG is noninvasive and has temporal resolution on the order of milliseconds, it is a useful tool for studying speech perception in adult and pediatric populations.

The EEG technique uses electrodes placed on the scalp of a listener to measure the electrical current from post-synaptic activity. To examine speech processing, an event-related paradigm is used and the EEG response is time-locked to auditory stimulus presentations. The EEG responses are then averaged to generate an auditory event-related potential (ERP) waveform (Figure 2).

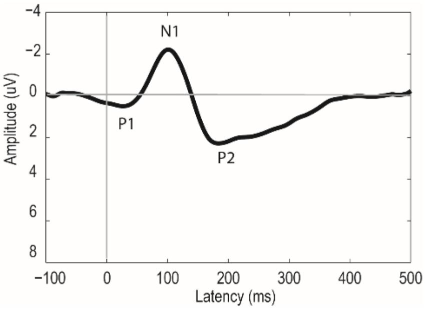


Figure 2. ERP waveform to an auditory stimulus showing the obligatory P1-N1-P2 response. Negative polarity plotted up.

The ERP waveform consists of a series of positive and negative peaks described by latency and amplitude values. Peak latency reflects the neural travel time through the auditory system and peak amplitude reflects the magnitude of the neural response to stimulus characteristics. Late auditory cortical potentials occur roughly 50 ms after stimulus onset, and the first positive and negative peaks of the waveform, the P1-N1-P2 complex, are obligatory because they can be recorded in the absence of attention. The P1-

N1-P2 response is commonly used to assess the neural coding of speech sounds and has been previously used to examine the neural coding of fricatives (Miller & Zhang, 2014; Tremblay, Billings, Friesen, & Souza, 2006). The P1 is the first positive peak in the sequence and occurs approximately 50ms after the stimulus. The P1 is thought to be generated by the primary auditory cortex, hippocampus, planum temporale, and lateral temporal regions (Key, Dove, & Maguire, 2005). The N1 is the first negative peak and occurs around 100 ms after the stimulus. The NI neural generators are thought to be bilateral primary and secondary auditory cortex (Naatanen et al., 1988). The P2 is the second positive peak and occurs approximately 180s after the stimulus. The P2 has many generators which include the primary and secondary auditory cortices and the reticular activating system (Key et al., 2005; Luck, 2005).

Speech Evoked Potentials

Previous studies have documented that the P1-N1-P2 components of the ERP response are sensitive to acoustic features of consonant and vowel speech sounds, making them suitable for examining neural coding of natural and synthetic speech sounds (Martin, Tremblay, & Korczak, 2008). Sharma, Marsh, and Dorman (2000) measured ERP responses elicited by synthetic /ba/-/pa/ and /ka/-ga/ contrasts differing in voice onset-time (VOT), the length of time between release of the consonant and the onset of voicing, and compared P1-N1-P2 responses across contrasts. The results indicated the voiced CV stimuli with shorter VOTs (VOTs between 0-30 ms), elicited N1 peak responses that were significantly earlier than the N1 responses elicited by the voiceless consonants with longer VOTs. The authors concluded that the N1 response reliably reflected the acoustic feature of VOT for voiced and voiceless bilabial and velar stop consonants.

Previous work has also examined whether consonant place of articulation differences can reliably be reflected in electrophysiological responses. Tavabi, Obleser, Dobel, & Pantev (2007) used MEG to examine whether the alveolar /d/ was differentially processed in the cortex relative to the velar /g/ with differing front-back vowel placements (/do/ /go/ /d∅/ /g∅/). Results howed an earlier and larger P1 peak response to the more frontal /d/ consonant than /g/. Furthermore, source localization results suggested the neural substrates differ for the different places of articulation, with frontal sounds such as /d/ activating deeper cortical areas than the back sound, /g/.

Agung, Purdy, McMahon, & Newall (2006) also previously recorded ERPs evoked by the naturally-produced phonemes /i/, /ɔ/, /m/, /a/, /u/, /s/, and /ʃ/ to determine whether different phoneme classes produced distinct ERP morphologies. Results revealed that the stimuli dominated by high frequency spectral energy, such as /s/ and /ʃ/, produced significantly smaller N1 and P2 amplitudes compared to stimuli dominated by lower frequencies. In addition, when they increased the duration of the stimuli, the longer stimuli produced smaller and later ERP peak amplitudes compared to the shorter duration stimuli. The authors concluded that ERPs are sensitive to spectral and temporal differences in naturally produced stimuli that cover the speech frequency range.

Neural Coding of Fricatives

Past research suggests that ERPs are sensitive to the acoustic characteristics of dynamically changing speech sounds. When fricative-vowel stimuli are used to elicit ERP responses, the response waveforms typically have multiple N1-P2 peak responses, reflecting the onset of the consonant and the onset of the vowel (Hari, 1991; Kaukoranta, Hari, & Lounasmaa, 1987). The double-peaked response elicited by a distinct change in the acoustic stimulus is typically referred to as the ‘acoustic change complex’ (ACC) (Martin & Boothroyd, 1999; Ostroff, Martin, & Boothroyd, 1998). These peaks to the vowel are denoted with a prime symbol, i.e. N1ˈ and P2ˈ.

Miller and Zhang (2014) previously used high density EEG to examine the P1-N1-P2 and ACC evoked by naturally produced fricative-vowel speech sounds in listeners with normal hearing. EEG data were collected using a 64-channel electrode montage, and ERP waveforms were elicited using /sɑ/ and /ʃɑ/ stimuli produced by a female talker. Results indicated that the P1-N1-P2 complex to the consonant and the ACC to the vowel significantly differed across stimuli, with N1 amplitudes being significantly larger for /sɑ/. The authors concluded that the spectral and dynamic formant transition cues that cue perception of fricatives are reliably coded in the auditory cortex. It remains unknown whether synthetic fricative stimuli would produce similar results or whether they are differentially processed by at the cortical level.

Neural Coding of Natural versus Synthetic Speech

Behavioral results indicate that natural and synthetic speech likely engage different cognitive mechanisms, and functional neuroimaging can potentially shed light on whether they engage different cortical structures. Functional neuroimaging studies have revealed that naturally produced phonetic segments activate multiple, overlapping cortical regions (Price, 2012). In general, fMRI and Positon Emission Tomography (PET) studies suggest during passive phonetic perception, the superior temporal lobe is activated bilaterally (Hickok & Poeppel, 2004). Some models of speech perception posit that cortical processing of phonemes then diverges into ventral and dorsal streams that are largely lateralized to the left hemisphere (Hickok & Poeppel, 2004). The ventral stream is thought to be involved in sound-to-meaning mapping and projects to superior temporal sulcus and to cortex in the posterior inferior temporal lobe. The dorsal stream is implicated in mapping sound to articulatory representations and projects toward parietal and frontal regions. Research suggests that cortical patterns of activation differ based on task demands (Price, 2012). It remains unclear whether synthetically produced phonemes activate similar areas of the cortex.

Some electrophysiological evidence exists that synthetic and natural speech could be processed differently in the auditory cortex, but previous ERP studies have mainly examined whether neural coding differs for synthetic versus natural vowels. Previous work by Swink and Stewart (2012) compared electrophysiological responses to natural and synthetic productions of the vowel /ɑ/. In the study, naturally produced stimuli were collected from both male and female talkers. Synthetic vowel tokens had a similar formant structure and had an equal duration to the naturally produced stimuli. EEG activity elicited by both the natural and synthetic vowels was recorded from 11 electrode sites, but only ERP waveform results from Cz were reported. The results indicated that peak P1, N1, and P2 latencies to the natural vowel were significantly earlier than those to the synthetic vowel. It remains untested whether fricative stimuli will show a similar pattern of results.

Specific Aims and Hypotheses

The specific aim of the present ERP study is to examine whether the synthetic and naturally produced fricatives /s/ and /ʃ/ are differentially coded in the auditory cortex at the phonetic level. Based on previous behavioral and electrophysiological data, we hypothesize that if different cognitive resources are used to process synthetic speech, ERP peak amplitude and latencies to the synthetic fricatives will be prolonged and smaller than those in response to natural speech. By using high density EEG measures, the present study also aims to examine whether natural and synthetic speech are differentially processed across left, midline, and right hemisphere sites. We hypothesize that synthetic speech will show less activation in the left electrode sites than natural speech. The collective results from this study will provide a better understanding of the brain mechanisms underlying the neural coding of natural and synthetic speech.

CHAPTER 2

METHODS

Subjects

Ten adults participated in the study (5 male and 5 female). Participants ranged in age from 19-27 years-old and were native speakers of American English. Subjects denied any history of speech, language, or neurological impairment. All subjects were right handed, per the Edinburgh Handedness Inventory (Oldfield, 1971), reported normal hearing sensitivity, and passed a hearing screening of a 1000Hz tone presented at 20dB HL. Informed consent for this study was obtained within compliance of the institutional human research protection program at The University of Minnesota (IRB 0804M31461).

Stimuli

The stimuli consisted of natural and synthetic consonant-vowel (CV) productions of the nonsense syllables /sɑ/ and /ʃɑ/. The vowel /ɑ/ was selected versus other vowel sounds because the combination of /s/ and /ʃ/ with vowel /ɑ/ results in a nonsense speech tokens. Controlling for lexical effects of EEG stimuli ensures that previously learned vocabulary would not affect cortical responses.   Each natural and synthetic stimulus had an exact duration of 350 ms. Peak latencies of evoked potential responses are sensitive to the acoustic parameters of stimuli, making strict control of duration imperative. For each stimulus, the fricative duration was 150ms and the vowel duration was 200ms.

Naturally Produced Speech Stimuli

The naturally-produced stimuli were edited using Sony Sound Forge 9.0 (Sony Creative Software). The tokens were recorded from an adult female who was a native speaker of American English in a sound booth (ETS-Lindgren Acoustic Systems). The talker produced the /sɑ/ and /ʃɑ/ syllables three times each into a high-fidelity microphone (Sennheiser), and the productions were digitally recorded to disk (44.1 kHz sampling rate). The best production of each stimulus was selected based on judgements from independent listeners that did not participate in the study. Once the stimuli were selected, the fricative and vowel durations were equated using temporal stretching and shrinking via the pitch synchronous overlap-add technique (Moulines & Charpentier, 1990). All stimuli were equated for root mean square (RMS) intensity level. Pilot testing suggested the digital processing of the stimuli did not affect the intelligibility of the syllables.

Synthetic Speech Stimuli

Synthetic /sɑ/ and /ʃɑ/ stimuli were created using HLSyn (Sensimetrics), HLSyn allows the user to control a small set of parameters that control a Klatt Synthesizer (Table 2).

Table 1. Summary of the acoustic parameters manipulated in HLSyn

HLsyn Parameter	Description
f1-f4	First four natural frequencies of vocal tract, assuming no local constrictions
f0	Fundamental frequency due to active adjustments of vocal folds
ab	Cross-sectional area of tongue blade constriction
ag	Average area of glottal opening between the membranous portion of the vocal fold
al	Cross-sectional area of constriction at the lips
an	Cross-sectional area of velopharyngeal port
ap	Area of the posterior glottal opening
dc	Change in vocal fold or wall compliances
ps	Subglottal pressure
ue	Rate of increase of vocal tract volume

Identical to the natural stimuli, the consonant portion of the synthetic stimuli was 150 ms and the vowel /ɑ/ was 200 ms in duration. The /s/ portion had a center frequency of 5000 Hz. The /ʃ/ portion had a center frequency of 2650 Hz. The /ɑ/ portion of each synthetic stimulus was identical. The F1 of /ɑ/ had a steady state frequency of 700 Hz. The F2 had a steady state frequency of 1200 Hz, and the F3 had a steady state value of 2700 Hz.

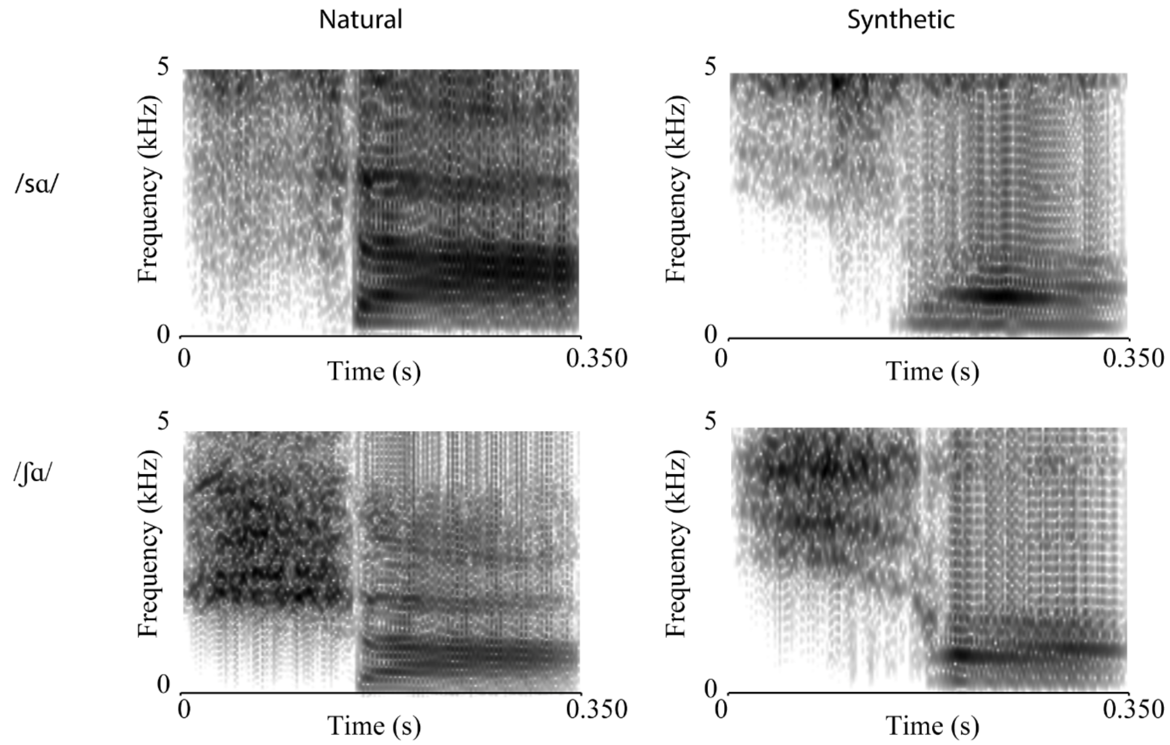


Figure 3. Spectrograms of the naturally produced and synthetic /sɑ/ and /ʃɑ/ stimuli used to elicit the ERP responses.

ERP Stimulus Presentation Protocol

For stimulus presentation, subjects were seated in a comfortable chair in an electrically and acoustically treated sound booth (ETS-Lindgren Acoustic Systems). Stimuli were presented in the sound field via bilateral loud speakers (M-Audio BX8a) located at approximately 60-degree azimuth angle to each subject. The stimuli were calibrated to 60 dB SPL relative to the subject's head before every session.

The natural and synthetic stimuli were presented to subjects in separate runs and presentation order of the runs was counterbalanced across subjects. Within each run, stimuli were presented using a passive listening, alternating short block design (Miller & Zhang, 2014; Zhang et al., 2011). Each block consisted of 20 stimuli in one category (20 tokens of /sɑ/), followed by a second block of 20 stimuli from the other category (20 tokens of /ʃɑ/). Blocks were alternated sequentially to ensure a sufficient and equal number of stimulus presentation from each category. The interstimulus interval between consecutive stimulus presentations in a block was randomized between 900-1000 ms to prevent adaptation. There was a 2 second silence periods between each block (Figure 4). To prevent a mismatch negativity response that might result from alternating block presentation, the first stimulus of each block was excluded from averaging (Figure 4).

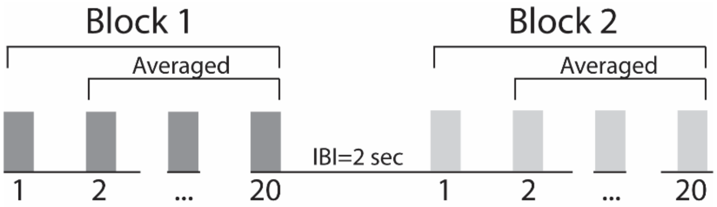


Figure 4. Illustration of the alternating block paradigm used to elicit ERP responses. 20 stimuli per block. Inter-stimulus interval was randomized between 900-1000ms. IBI indicates inter-block interval. The first stimulus of each block was not included in the averages to avoid a MMN.

EEG Data Acquisition

EEG activity was recorded using the Advanced Neuro Technology EEG system and a 64 channel Waveguard Cap (ANT, Inc.,) (Rao, Zhang, & Miller, 2010). Continuous EEG data were band pass filtered from 0.016 to 200 Hz and digitized using a 512 Hz sampling rate. The Ag/AgCl electrodes were sewn into the cap using the international 10-20 montage and intermediate locations. The ground electrode was located at the AFz position. The average electrode impedance was kept below 5k Ohms throughout the experiment. During the EEG recording, subjects viewed a muted, subtitled movie of their choice on a 20-inch LCD TV located 2.5 meters in front of the listener. Subjects were instructed to ignore the stimuli and attend to the movie. The entire experimental session lasted approximately 60 minutes.

ERP Waveform Analysis

Analysis of the averaged ERP waveforms from individual subjects was completed offline using the Advanced Neuro Technology EEG system (Advanced Source Analysis version 4.7) and MATLAB (Mathworks). The raw EEG data were bandpass filtered from 0.5-40 Hz. The ERP epoch was 800 ms and consisted of a 100ms prestimulus baseline followed by a 700 ms recording window. The artifact rejection criterion for individual trials was set to +/- 50 uV. After averaging, 112 trials remained for the different stimulus conditions. Linked mastoids was used as the reference for the offline ERP waveform analysis.

Peak amplitude and latency of the P1-N1-P2 complex elicited by the fricative and the N1ˈ and P2ˈ elicited by the vowel of the stimuli were extracted from the averaged ERP waveforms for each subject. Based on the grand average waveforms, the following latency ranges were used to extract P1-N1-P2 peaks to the fricative: P1 35 to 80ms; N1 85 to 170ms P2 165 to 245ms. and N1’-P2’ peaks to the vowel ACC peaks to the CV transition and vowel latency: N1ˈ; 240 to 310ms, P2ˈ 300 to 380ms.

Statistical Analysis

Effects of *speech condition* (naturally produced and synthetic) and *phonetic identity* (/s/ and /ʃ/) on peak ERP waveform amplitudes and latencies from individual subject data were assessed using a repeated-measures analysis of variance (R-ANOVA) in Systat (Version 13.1). Because auditory ERP responses are typically largest at central electrode sites (Luck, 2003), the central electrodes were grouped for analysis to examine hemisphere effects on peak amplitudes and latencies, and *laterality* (left, middle, right hemisphere electrode sites) was also included as within-subject factors in the ANOVA. The left central electrodes included T7, TP7, C3, C5, CP3, CP5 and electrodes TP8, C4, C6, CP4, and CP6 on the right hemisphere. Midline central electrodes included C1, Cz, C2, CP1, CPz, and CP2 (Figure 5).

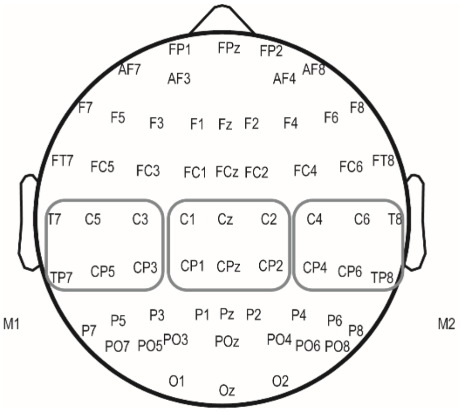


Figure 5. Left, midline, and right central electrode groupings used in the statistical analysis.

CHAPTER 3

RESULTS

ERP Results

Clear P1-N1-P2 responses to the fricative and N1ˈ, P2ˈ to the vowel were observed across all electrode regions for the natural and synthetic /sɑ/ (Figure 6) and /ʃɑ/ stimuli (Figure 7). Grand mean peak amplitude, peak latency, and standard deviations used in the statistical analysis for each ERP component of interest are summarized in Table 2. Separate repeated-measures R-ANOVAs for P1, N1, P2, N1ˈ, and P2ˈ peak latencies and amplitudes were performed. Table 3 summarizes the full model R-ANOVA results for each component.

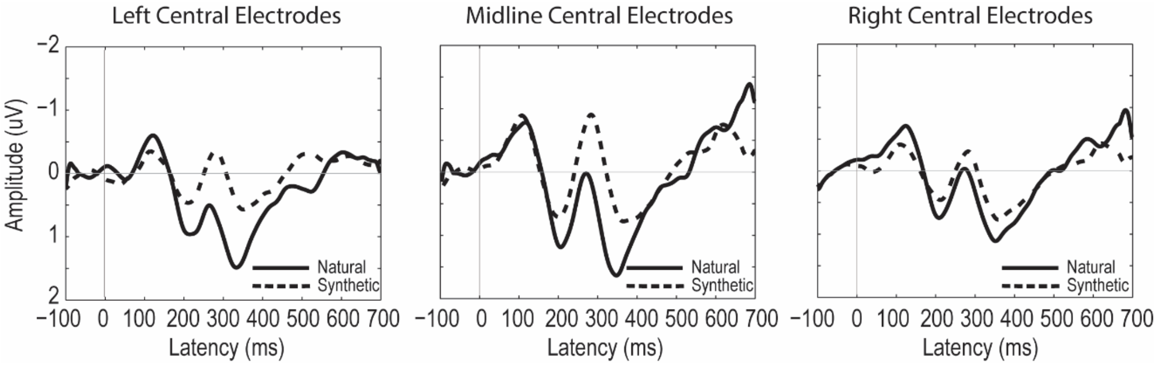


Figure 6. Grand mean ERP waveforms for natural and synthetic /sɑ/ stimuli for the left, midline, and right central electrode groups. Linked mastoid reference. Negative polarity plotted up.

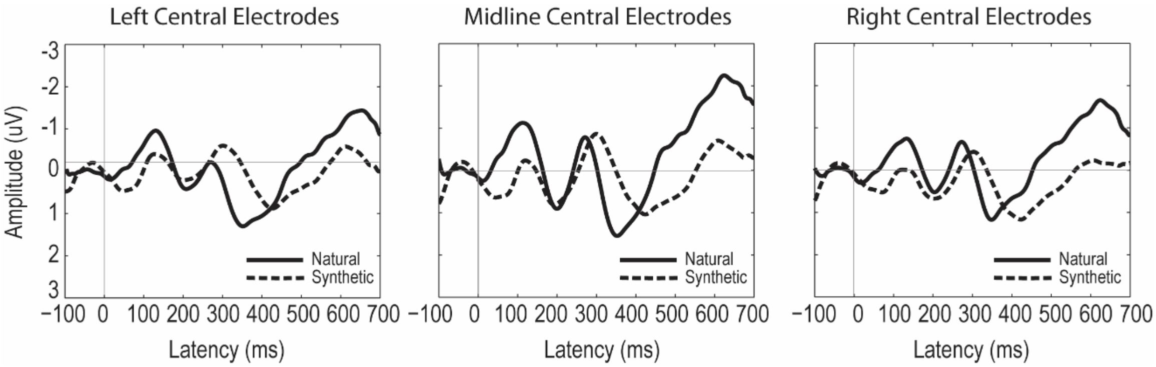
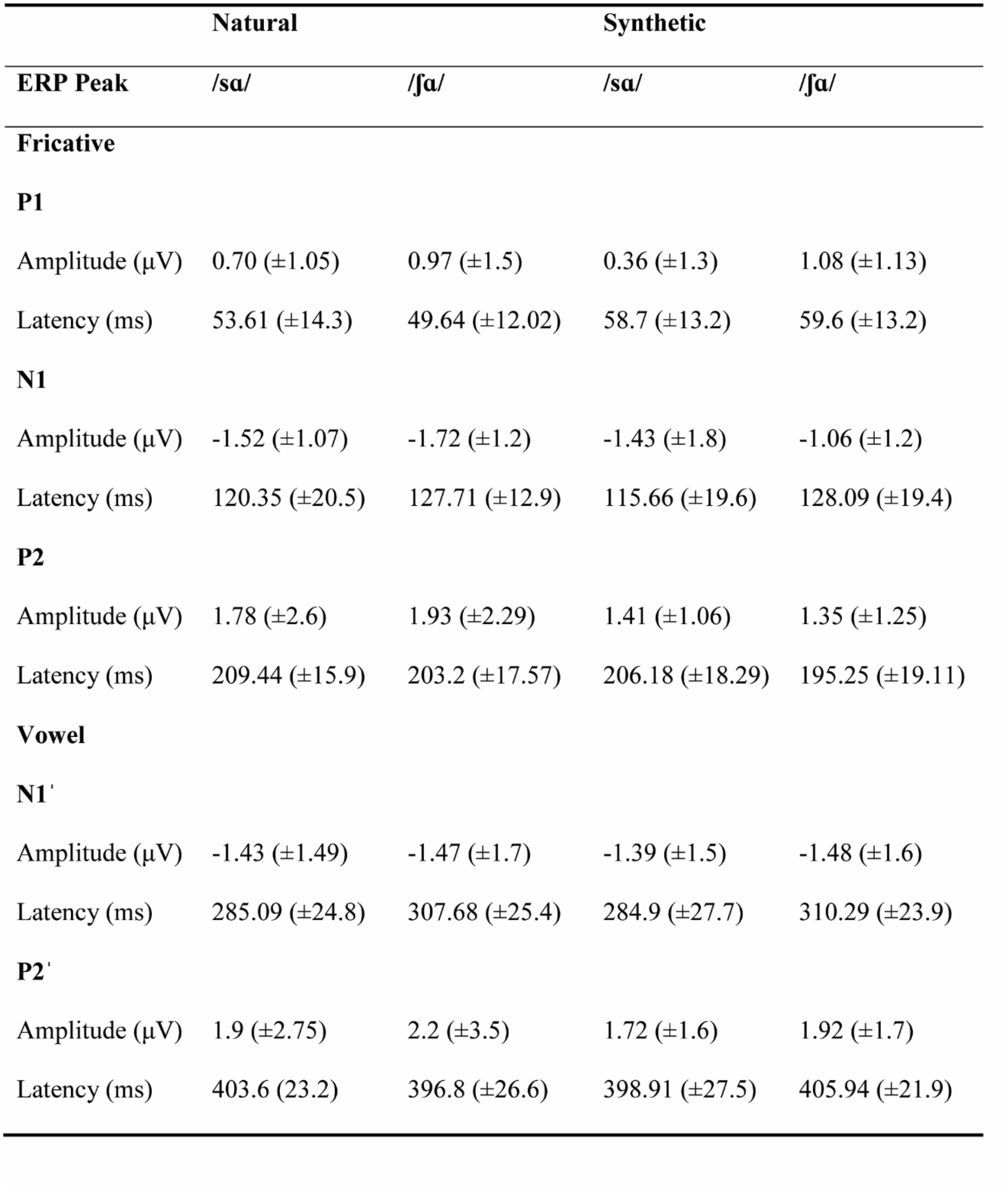


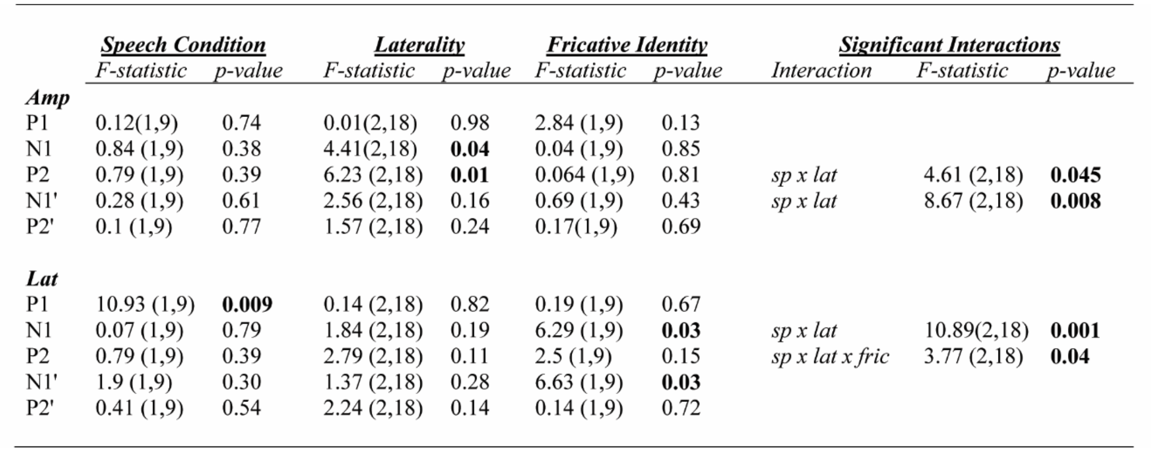
Figure 7. Grand mean ERP waveforms for natural and /ʃɑ/ stimuli for the left, midline, and right central electrode groups. Linked mastoid reference. Negative polarity plotted up.

Table 2. Peak amplitude and latency values (± 1 standard deviation) averaged across the left central, midline central, and right central electrode groups) for the P1, N1, P2, N1’ and P2’ components used in the statistical analysis.



	Natural		**Synthetic**	
ERP Peak	**/sɑ/**	**/ʃɑ/**	**/sɑ/**	**/ʃɑ/**
Fricative				
P1				
Amplitude (μV)	0.70 (±1.05)	0.97 (±1.5)	0.36 (±1.3)	1.08 (±1.13)
Latency (ms)	53.61 (±14.3)	49.64 (±12.02)	58.7 (±13.2)	59.6 (±13.2)
N1				
Amplitude (μV)	-1.52 (±1.07)	-1.72 (±1.2)	-1.43 (±1.8)	-1.06 (±1.2)
Latency (ms)	120.35 (±20.5)	127.71 (±12.9)	115.66 (±19.6)	128.09 (±19.4)
P2				
Amplitude (μV)	1.78 (±2.6)	1.93 (±2.29)	1.41 (±1.06)	1.35 (±1.25)
Latency (ms)	209.44 (±15.9)	203.2 (±17.57)	206.18 (±18.29)	195.25 (±19.11)
Vowel				
N1'				
Amplitude (μV)	-1.43 (±1.49)	-1.47 (±1.7)	-1.39 (±1.5)	-1.48 (±1.6)
Latency (ms)	285.09 (±24.8)	307.68 (±25.4)	284.9 (±27.7)	310.29 (±23.9)
P2'				
Amplitude (μV)	1.9 (±2.75)	2.2 (±3.5)	1.72 (±1.6)	1.92 (±1.7)
Latency (ms)	403.6 (23.2)	396.8 (±26.6)	398.91 (±27.5)	405.94 (±21.9)

Table 3. Repeated-measures ANOVA results summary for peak amplitudes (Amp) and latencies (Lat). Within-subjects main effects of speech condition (natural, synthetic), laterality (left, midline, right hemisphere electrodes), and fricative identity (/sɑ/, /ʃɑ/), were included in the analysis. Significant main effects are indicated in bold (*p*<0.05). All significant interactions observed between the within-subject factors are listed in the far right column (sp=speech condition; lat=laterality; fric=fricative identity.)



	Speech Condition		***Laterality***		***Fricative Identity***		***Significant Interactions***		
	F-statistic	*p-value*	*F-statistic*	*p-value*	*F-statistic*	*p-value*	*Interaction*	*F-statistic*	*p-value*
Amp									
P1	0.12(1,9)	0.74	0.01(2,18)	0.98	2.84 (1,9)	0.13			
N1	0.84 (1,9)	0.38	4.41(2,18)	**0.04**	0.04 (1,9)	0.85			
P2	0.79 (1,9)	0.39	6.23 (2,18)	**0.01**	0.064 (1,9)	0.81	*sp x lat*	4.61 (2,18)	**0.045**
N1'	0.28 (1,9)	0.61	2.56 (2,18)	0.16	0.69 (1,9)	0.43	*sp x lat*	8.67 (2,18)	**0.008**
P2'	0.1 (1,9)	0.77	1.57 (2,18)	0.24	0.17(1,9)	0.69			
Lat									
P1	10.93 (1,9)	**0.009**	0.14 (2,18)	0.82	0.19 (1,9)	0.67			
N1	0.07 (1,9)	0.79	1.84 (2,18)	0.19	6.29 (1,9)	**0.03**	*sp x lat*	10.89(2,18)	**0.001**
P2	0.79 (1,9)	0.39	2.79 (2,18)	0.11	2.5 (1,9)	0.15	*sp x lat x fric*	3.77 (2,18)	**0.04**
N1'	1.9 (1,9)	0.30	1.37 (2,18)	0.28	6.63 (1,9)	**0.03**			
P2'	0.41 (1,9)	0.54	2.24 (2,18)	0.14	0.14 (1,9)	0.72			

P1 Results

Peak P1 latencies evoked by naturally produced fricatives were significantly earlier than those evoked by synthetic speech [F(1,9)=10.932, *p*=0.009]. The main effects of *laterality* (left, midline, and right) and *fricative identity* (/sɑ/, /ʃɑ/) were not significant (*p*>0.05). All interactions between main effects were also non-significant for P1 latencies (*p*>0.05). For P1 amplitudes, the main effects of *speech condition* (natural, synthetic), *laterality, and fricative identity* and all interactions between main effects were not significant (*p* > 0.05).

N1 Results

Repeated-measures ANOVA revealed /sɑ/ elicited significantly earlier N1 peak latencies than /ʃɑ/ in both the natural and synthetic speech conditions [F(1,9)=6.289, *p*=0.03]. The two-way interaction between *speech condition* x *laterality* was also significant [F(2,18) = 10.89, *p*=0.001]. A one-way post-hoc ANOVA indicated that for natural speech, N1 latencies significantly differed across left, midline, and right hemisphere sites [F(2,18)=6.2, *p*=0.013]. Post-hoc paired comparisons indicated that N1 latencies were significantly earlier for natural speech at the midline electrodes relative to the right hemisphere electrodes (*p*=0.049). For synthetic speech, there were no significant differences in N1 latency for the left, midline, or right central electrode sites (*p*>0.05).

Repeated measures ANOVA revealed a significant main effect of *laterality* for N1 amplitudes for both natural and synthetic speech [F(2,18)=4.4, *p*=0.035], but post-hoc paired comparisons indicated that differences across the three levels (left vs. midline; left vs. right; and midline vs. right) did not significantly differ (*p*>0.05).

P2 results

Repeated measures ANOVA for P2 latencies revealed a significant three-way interaction between *speech condition*, *laterality,* and *fricative identity* [F(2,18)=3.77, *p*=0.04]. Post-hoc analysis indicated that P2 latencies for naturally produced /sɑ/ and /ʃɑ/ stimuli were differentially coded across the three hemisphere sites as indicated by the significant *fricative identity* x *laterality* interaction for natural speech [F(2,18)=3.6, *p*=0.049]. P2 latencies for synthetic fricatives did not significantly differ across the three hemisphere sites [F(2,18)=2.6, *p*>0.05].

Repeated measures ANOVA for P2 amplitudes indicated a significant main effect of *laterality* [F(2,18) = 6.229, *p*=0.009]. There was also a significant *speech condition* x *laterality* interaction [F(2,18) = 4.6, *p* = 0.045]. Post-hoc analysis of the significant interaction suggested that P2 amplitudes for naturally produced fricatives at left and midline sites did not significantly differ (*p*>0.05), but synthetic fricatives produced significantly larger P2 amplitudes at midline sites compared to left hemisphere sites (*p*=0.004).

N1ˈ and P2ˈ (ACC) Results

Repeated-measures ANOVA indicated that N1ˈ, the first negative peak of the ACC to the vowel, was significantly earlier for /sɑ/ than /ʃɑ/ for natural and synthetic speech [F(1,9)=6.626, *p*=0.03]. For N1ˈ amplitudes, there was a significant interaction between *speech condition* and *laterality* [F(2,18)=8.67, *p*=0.008]. Within natural speech, the effect of laterality approached significance [F(2,18)=3.241, *p*=0.06]. For synthetic speech, the effect of laterality was not significant [F(2,18)=2.7, *p*=0.124]. P2ˈ peak analysis indicated there were no significant main effects or interactions between main effects for P2ˈ latencies or amplitudes.

CHAPTER 4

DISCUSSION

The present study examined whether the neural coding of the sibilant /sɑ/-/ʃɑ/ contrast differed for natural versus synthetic productions. Based on previous behavioral research, we hypothesized that cortical responses would be more robust and efficient for natural speech. Consistent with our hypothesis, P1 cortical responses were significantly earlier for natural versus synthetic fricatives. In addition, naturally produced fricatives showed significant hemisphere site effects for the P1-N1-P2 complex. In contrast, synthetic fricatives were processed similarly across the left, midline, and right hemisphere sites. Finally, the hemisphere site effects for natural versus synthetic speech were also observed for the following vowel. In total, the results of the present study suggest fricative speech stimuli are differentially processed in the auditory cortex depending on if they are naturally or synthetically produced. The clinical implications of the study for the communication disorders field and comparisons to previous behavioral and electrophysiological results will be discussed.

Natural versus Synthetic Speech

The finding that P1 latency was significantly earlier for natural compared to synthetic speech suggests that differences in neural coding emerge at an early, pre-attentive level. This early cortical difference at P1 coupled with the hemisphere site effects observed for N1, P2, and N1ˈ components suggests that natural speech activates different cortical processing pathways compared to synthetic speech. As reviewed.

previously, speech perception occurs when the acoustically variable signal is mapped onto abstract phonological representations in auditory cortex, and neurophysiological studies suggest this mapping process likely occurs in a series of multiple, hierarchical stages (Hickok & Poeppel, 2015). Both speech and non-speech sounds are thought to activate superior temporal gyrus bilaterally and that left lateralized activation for speech arises in later processing stages (Hickok & Poeppel, 2015). It is possible the pattern of results observed in the present study indicate that synthetic productions are processed more like non-speech sounds, where there is an absence of later, left-dominant activation. This view is supported by the findings of Rinne and colleagues (1999) who used high density EEG and measured cortical responses elicited by sounds on a continuum from non-speech (tones) to speech (vowels). They found that as the stimuli became more speech-like, left temporal activation systematically increased. The lack of hemisphere effects for the P1-N1-P2 peaks to synthetic speech in the present study could indicate that the stimuli were processed more acoustically at all levels of cortical processing.

Synthetic Speech in the Speech-Language Pathology Domain

The differential activation of auditory cortex in response to natural and synthetic speech has important clinical implications for the speech-language pathology field. Augmentative and alternative communication (AAC) aids and devices are used extensively in the speech-language pathology domain and allow persons with speech and language impairments to communicate more effectively. Speech generating devices (SGDs) for verbal communication primarily use synthetic speech in order to maximize the number of unique utterances that can be produced. The synthesized speech technology in SGDs has improved in recent years and transformed from robotic speech to an array of natural-sounding male, female, and child-like voices (Beukelman, Mirenda, & Beukelman, 2013). Currently, there are three main types of synthesized speech used in SGD and AAC devices. Text-to-speech synthesizers are the most common and generate speech by coding text that is stored within the AAC device into corresponding phonemes, and then converting the digital signals into acoustic waveforms. Text-to-speech synthesizers do not store speech in a digital form, per se, instead they create synthesized speech based on a mathematical algorithm revolving around rule-generated speech. A second type of text-to-speech synthesizer uses diphone-based strategies to produce speech. Diphones are extracted from carrier words produced by human talkers resulting in a more natural sounding product than those from traditional text-to-speech synthesizers. Finally, AAC devices can use digitized speech. Digitized speech is a form of electronic speech produced primarily from natural speech recorded to disk (Beukelman et al., 2013). Previous behavioral studies have examined whether the new synthesized speech technologies used in AAC devices are as intelligible as natural productions.

In an early study, Koul and Allen (1993) examined whether intelligibility of natural versus synthetic speech used in AAC devices differed when presented in background noise. CVC words were presented to adult listeners in three forms: DecTalk Paul (male), DecTalk Betty (female), and natural speech (adult male). Lists of words in twelve-talker babble were presented at three different signal-to-noise ratios (SNRs): 0 dB, 15 dB, and 25 dB. Percent correct intelligibility scores for each type of speech were computed at each SNR. Results suggested that intelligibility scores were significantly higher for natural speech than either of the two types of synthetic speech across SNRs.

Error pattern analysis indicated that scores for natural speech were significantly more intelligible than the two types of synthetic speech. The breakdown of specific phoneme errors showed initial errors in synthetic speech stimuli occurred primarily for nasals, stops, and fricatives across all three SNRs. For the synthesized speech, nasals, stops and the voiceless fricative /s/ accounted for the most errors in the phoneme-final position. In the phoneme initial position, nasals and stops accounted for the majority of errors for synthetic speech. For natural speech in both the phoneme initial and final positions, the largest number of errors occurred for fricatives, nasals, and stops. The DecTalk synthesizer is commonly used in AAC devices, and the results of the Koul and Allen (1993), in conjunction with the results of the present study, suggest that there may be a disadvantage to using this output form in AAC devices, especially when in a classroom or noisy environment.

In a more recent study, Pinkowski-Ball, Reichle, and Munson (2012) examined the intelligibility of speech produced by a variety of new AAC technologies in preschool-aged children in typical noise environments. Single words were presented using natural speech, and two types of synthetic speech: AT&T voice Michael and DECTalk voice Paul. Intelligibility was scored as the mean percentage of words repeated correctly. Results showed the average intelligibility for human speech was 97.5%, AT&T Michael was 91.4%, and DECtalk Paul was 84.75%. DECtalk is still a leading synthesizer used in AAC devices and voice Paul has previously been shown to be the most intelligible of the DECtalk voice options (Pinkoski-Ball et al., 2012). The results of this study demonstrate that when comparing different speech outputs in a realistic setting (classroom and school hallway), the most commonly used speech synthesizer is the least intelligible.

Furthermore, for this population, natural speech was the most intelligible which is of critical importance because young students using AAC devices are still acquiring language.

In addition to AAC devices, speech-language pathologists also work with patients that produce alaryngeal speeah, speech produced without the larynx. Similar to work examining differences between natural and synthetic speech, Evitts and Searl (2006) compared whether alaryngeal speech requires greater cognitive resources to process relative to normal laryngeal speech and synthetic speech. The authors examined behavioral reaction times for single words produced naturally, three types of alaryngeal speech (electrolaryngeal speech, esophageal speech and tracheoesphogeal speech), and synthetic speech. To control for differences in the duration of stimuli, response reaction time to the stimuli was compared to the mean stimulus duration, and a ratio representing cognitive processing load was computed for each subject. The results indicated that alaryngeal speech required significantly more cognitive processing effort than naturally produced speech. Of note, of the three classes of material, synthetic speech required the greatest cognitive processing demands, meaning it was more difficult to process than even highly unnatural, alaryngeal speech. The authors concluded that differences in processing demands suggest that synthetic speech is entirely different than speech produced by a human, even speech from an electrolarynx.

Although prevalent, the use of synthetic speech is not limited to the field of speech-language pathology and AAC devices. Synthetic speech is heard commonly in everyday life via ATM’s, cell phone voice command systems, and GPS navigation systems, to name a few. In a recent study, Wolters et. al (2015) investigated the use of synthetic speech to remind older adults to take their medications. The multi-dimensional study assessed whether older adults, with a range of hearing from normal to some age-related hearing loss, had a more difficult time recalling medication reminders when they were presented with synthetic speech outputs as opposed to a natural human voice. When presented with known medications, participants had similar recall rates for all types of speech output. However, when presented with unknown medications, the recall rates were much lower when recalling synthetic speech stimuli (52.2% accuracy) than natural speech (64.8% accuracy). The study concluded that synthetic speech can be a useful tool for medication reminders, but it is potentially dangerous to rely on it as a sole teaching method, especially for new or unfamiliar medications. The best approach is a multimodal approach including repetition of medications, explanations of medications, and familiarity gained from a human voice before relying on solely synthetic speech.

There is ample evidence from the AAC and alaryngeal speech literature that synthetic speech is less intelligible than natural speech (Evitts & Searl, 2006; Pinkoski-Ball et al., 2012), is more susceptible to degradation from noise (Koul & Allen, 1993), and requires greater cognitive processing resources than natural speech (Pisoni, 1981). The data from the present study support the notion that these behavioral results likely reflect the different cortical circuits activated by natural and synthetic speech. The present electrophysiological results might indicate that natural speech should be used whenever possible.

Limitations and Future Directions

The present study only examined the neural coding of one synthetic and naturally produced fricative-vowel contrast. While synthetic fricatives are subject to the most misperception compared to other classes of speech sounds, making it an important class to study, it is possible that results would not generalize to other speech sounds. Thus, future EEG studies should examine whether other consonant classes show the similar pattern of cortical activations for natural and synthetic speech. Future studies should also examine whether the same pattern of results would be observed for fricatives in other phonological contexts, i.e. vowel-consonant positions. It might be that onset coding of fricatives requires different cortical mechanisms than fricatives in the coda position.

Another limitation of the present study is that only electrophysiological measures to brief fricative-vowel stimuli were collected. While pilot studies showed the natural and synthetic speech stimuli were equally intelligible (Miller & Zhang, 2014), it remains unknown whether the differences across stimuli would predict other ecologically valid measures of behavioral speech perception. Future studies should examine whether ERP peak measures for the synthetic and natural speech predict behavioral word and sentence performance in a variety of listening situations.

The spatial resolution of EEG is limited compared to other imaging techniques, so the current results would be strengthened if the hemispheric differences were also observed using fMRI or MEG measures. The use of fMRI would enhance our ability to make specific claims about what cortical structures are involved in the coding of natural and synthetic speech. The use of EEG only allows us to conclude that there were differences across natural and synthetic stimuli.

Finally, in the present study, differences in natural and synthetic speech were seen at P1, the earliest response from auditory cortex. Auditory P1 is known to be a sensitive neural marker of sensory gating (Korzyukov et al., 2007), the reduction in peak ERP amplitudes with repeated stimulus presentation. Sensory gating is thought to originate from the cortical-thalamic loop which acts as a gate to prevent auditory cortex from being flood with extraneous information (Korzyukov et al., 2007). It would be interesting to examine whether synthetic and naturally produced speech are differentially gated by listeners. It remains possible that the differences in ERP amplitude between natural and synthetic speech found in this study result from synthetic speech being gated to a greater degree.

Overall, the results of the present study suggest that neural coding differs for natural and synthetic speech in adults with normal hearing, and the differences in processing occurs at the earliest levels of cortical processing. Whether the same pattern of results emerges for persons with communication disorders such as hearing loss, autism, or aphasia remains to be determined.

REFERENCES

Agung, K., Purdy, S. C., McMahon, C. M., & Newall, P. (2006). The use of cortical auditory evoked potentials to evaluate neural encoding of speech sounds in adults. *J Am Acad Audiol,* 17(8), 559-572.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Sci Am,* 225(2), 82-90.

Beukelman, D. R., Mirenda, P., & Beukelman, D. R. (2013). *Augmentative and alternative communication : supporting children and adults with complex communication needs* (4th ed.). Baltimore: Paul H. Brookes Pub.

Borden, G. J., Harris, K. S., & Raphael, L. J. (2003). *Speech science primer : physiology, acoustics, and perception of speech* (4th ed.). Philadelphia: Lippincott Williams & Wilkins.

Borden Gloria J, H. K. S., Raphael Lawrence K. (2011). *Speech Science Primer: physiology, acoustics, and perception of speech* Lippincott Williams & Wilkins

Clark, J. E. (1983). Intelligibility comparisons for two synthetic and one natural speech source. *Journal of Phonetics,* 11, 37-49.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Lang Speech,* 40 (Pt 2), 141-201.

Evitts, P. M., & Searl, J. (2006). Reaction times of normal listeners to laryngeal, alaryngeal, and synthetic speech. *J Speech Lang Hear Res,* 49(6), 1380-1390.

Fant, G. (1960). *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage,: Mouton.

Gazzaniga, M., & Mangun, G. (2014). *The cognitive neurosciences* (Firth edition. ed.). Cambridge, Massachusetts: The MIT Press.

Gazzaniga, M. S. (2009). *The Cognitive Neurosciences*. Cambridge, Massachusetts The MIT Press.

Greene, B. G. P., D.B.; Nusbaum, H.C. (2005). Perception of synthetic speech generated by rule. *Proceedings of the IEEE,* 73(11).

Hari, R. (1991). Activation of the human auditory cortex by speech sounds. *Acta Otolaryngol Suppl,* 491, 132-137; discussion 138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition,* 92(1-2), 67-99.

Hickok, G., & Poeppel, D. (2015). Neural basis of speech perception. *Handb Clin Neurol,* 129, 149-160.

Johnson, K. (2003). *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishing.

Kaukoranta, E., Hari, R., & Lounasmaa, O. V. (1987). Responses of the human auditory cortex to vowel onset after fricative consonants. *Exp Brain Res,* 69(1), 19-23.

Key, A. P., Dove, G. O., & Maguire, M. J. (2005). Linking brainwaves to the brain: an ERP primer. *Dev Neuropsychol,* 27(2), 183-215.

Korzyukov, O., Pflieger, M. E., Wagner, M., Bowyer, S. M., Rosburg, T., Sundaresan, K., Elger, C. E., & Boutros, N. N. (2007). Generators of the intracranial P50 response in auditory sensory gating. *Neuroimage,* 35(2), 814-826.

Koul, R. K., & Allen, G. D. (1993). Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in presence of noise. *J Speech Hear Res,* 36(4), 790-798.

Ladefoged, P. (1962). *Elements of Acoustic Phonetics*. Chicago, IL: University of Chicago Press.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Hum Factors,* 25(1), 17-32.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, Mass.: MIT Press.

Martin, B. A., & Boothroyd, A. (1999). Cortical, auditory, event-related potentials in response to periodic and aperiodic stimuli with the same spectral envelope. *Ear Hear,* 20(1), 33-44.

Martin, B. A., Tremblay, K. L., & Korczak, P. (2008). Speech evoked potentials: from the laboratory to the clinic. *Ear Hear,* 29(3), 285-313.

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

Miller, S., & Zhang, Y. (2014). Neural coding of phonemic fricative contrast with and without hearing aid. *Ear Hear,* 35(4), e122-133.

Moulines, E., & Charpentier, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Commun,* 9.

Naatanen, R., Sams, M., Alho, K., Paavilainen, P., Reinikainen, K., & Sokolov, E. N. (1988). Frequency and location specificity of the human vertex N1 wave. *Electroencephalogr Clin Neurophysiol,* 69(6), 523-531.

Nusbaum, H., Dedina, J. J., & Pisoni, D. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. *Speech Research Laboratory Technical Note,* 84(2).

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychol Rev,* 85(3), 172-191.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia,* 9, 97-113.

Ostroff, J. M., Martin, B. A., & Boothroyd, A. (1998). Cortical evoked response to acoustic change within a syllable. *Ear Hear,* 19(4), 290-297.

Pinkoski-Ball, C. L., Reichle, J., & Munson, B. (2012). Synthesized speech intelligibility and early preschool-age children: comparing accuracy for single-word repetition with repeated exposure. *Am J Speech Lang Pathol,* 21(4), 293-301.

Pisoni, D. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America,* S98.

Pisoni, D., & Levi, S. V. (2007). Representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage,* 62(2), 816-847.

Rao, A., Zhang, Y., & Miller, S. (2010). Selective listening of concurrent auditory stimuli: an event-related potential study. *Hear Res,* 268(1-2), 123-132.

Rinne, T., Alho, K., Alku, P., Holi, M., Sinkkonen, J., Virtanen, J., Bertrand, O., & Naatanen, R. (1999). Analysis of speech sounds is left-hemisphere predominant at 100-150ms after sound onset. *Neuroreport,* 10(5), 1113-1117.

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *J Acoust Soc Am,* 108(6), 3030-3035.

Slowiaczek, L. M., & Pisoni, D. (1982). Effects of practice on a speeded classification of natural and syntheic speech. *Research on Speech Perception Progress Report,* 7, 701-712.

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

Swink, S., & Stuart, A. (2012). Auditory long latency responses to tonal and speech stimuli. *J Speech Lang Hear Res,* 55(2), 447-459.

Tavabi, K., Obleser, J., Dobel, C., & Pantev, C. (2007). Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. *Eur J Neurosci,* 25(10), 3155-3162.

Tremblay, K. L., Billings, C. J., Friesen, L. M., & Souza, P. E. (2006). Neural representation of amplified speech sounds. *Ear Hear,* 27(2), 93-103.

Wolters, M. K., Johnson, C., Campbell, P. E., DePlacido, C. G., & McKinstry, B. (2015). Can older people remember medication reminders presented using synthetic speech? *J Am Med Inform Assoc,* 22(1), 35-42.

Zeng, F. G., & Turner, C. W. (1990). Recognition of voiceless fricatives by normal and hearing-impaired subjects. *J Speech Hear Res,* 33(3), 440-449.

Zhang, Y., Koerner, T., Miller, S., Grice-Patil, Z., Svec, A., Akbari, D., Tusler, L., & Carney, E. (2011). Neural coding of formant-exaggerated speech in the infant brain. *Dev Sci,* 14(3), 566-581.

APPENDIX: ABBREVIATIONS

AAC	augmentative and alternative communication
ACC	acoustic change complex
ATM	automated teller machine
CV	consonant-vowel
CVC	consonant-vowel-consonant
dB	decibel
EEG	electroencephalography
ERP	auditory event related potential
fMRI	functional magnetic resonance imaging
GPS	global positioning system
HL	hearing loss
Hz	hertz
MEG	magnetoencephalography
MMN	mismatched negativity
ms	milliseconds
PET	position emission tomography
SGD	speech generating device
RMS	root mean square
SNR	signal-to-noise ratios
VOT	voice-onset time

CURRICULUM VITAE

Allison Brown, B.S.
425 Bauer Ave
Louisville, KY 40207
(859) 576-9930 (cell phone)
ajbrow12@louisville.edu

EDUCATION

University of Louisville, Louisville, KY

M.S., Communicative Disorders 2015-present

University of Kentucky, Lexington, KY

B.S., Communication Sciences and Disorders 2012-2015

RESEARCH EXPERIENCE

University of Louisville, Louisville, KY
Department of Otolaryngology Head and Neck Surgery and Communicative Disorders
Research mentor: Dr. Sharon Miller

- Master's thesis project entitled *Neural Coding of Natural Versus Synthetic Speech*
    - Auditory event-related potential data collection and analysis

Kent College of Social Work
Research mentor: Dr. Andy Frey

- Elementary First Steps Next and HomeBase Behavior Intervention study
- Preschool First Steps Next Behavior Intervention study
    - Parent screenings
    - Parental consent
    - Baseline data collection
    - Pre-K Standardized Behavior Assessment

University of Kentucky, Lexington, KY
College of Health Sciences
Research mentor: Dr. Daniel Croake 2014-2015

- Assessment of relationship between diaphragm, laryngeal function, and articulation
- Comparison to vocal function of normal voice versus vocal fold paralysis
    - Data collection
    - Data entry
    - Subject recruitment

College of Health Sciences
Research mentor: Dr. Jane Kleinert 2015

- AAC Educational Modules for Teachers
    - Viewed AAC educational modules
    - AAC educational modules exams
    - Reviewed and critiqued AAC educational modules

College of Education
Research mentor: Dr. Allan Allday 2014

- Middle School Aged Behavior Intervention
    - Data Collection

PRESENTATIONS

Posters

Regional meetings

Brown, A. and Miller, S.E. (2017). Neural coding of natural and synthetic speech. Poster presentation at the Kentucky Speech-Language-Hearing Association annual meeting, February 23, Lexington, KY.

*Poster awarded 1st place in the student research symposium.

PROFESSIONAL MEMBERSHIPS AND ACTIVITIES

National Student Speech Language Hearing Association	2013-present
Kentucky Speech Language Hearing Association	2013-present