UNIVERSITY
OF TURKU

Auditory Activity Evoked by Self-Produced Foreign Phonemes Changes as Pronunciation Improves

Master’s Degree Program in Human Neuroscience
Faculty of Medicine
Master's thesis

Author:
Anni Varjonen

Supervisor:
Henry Railo, PhD

May 2021

Master's thesis

Subject: Phoneme learning and SIS
Author(s): Anni Varjonen
Title: Auditory Activity Evoked by Self-Produced Foreign Phonemes Changes as Pronunciation Improves
Supervisor(s): Dr. Henry Railo
Number of pages: 41
Date: 25.05.2021

Abstract

Phoneme learning is a complex process that involves the integration of auditory perception and motor activity, and this phenomenon is a central concept in our ability to produce coherent speech. Although phoneme learning has been studied using event-related potentials (ERPs) in the past, most of the research has focused on listening paradigms. Little research has been done on electroencephalogram (EEG) correlates that take place during the active pronunciation of a foreign phoneme. Our study addressed this gap in literature by focusing on an ERP amplitude difference called Speaking Induced Suppression (SIS), during the pronunciation of an unfamiliar phoneme. The SIS event refers to the brain's tendency to show suppressed auditory responses to self-produced speech in comparison to the same sounds that are passively heard (Niziolek et al., 2013). SIS is thought to reflect a process in the speech production system that compares how well produced speech matches the intended speech (Guenther & Vladusich, 2011), and there seems to be more suppression in the auditory cortex when the produced and attempted sound match closely (Ventura et al., 2009). Our study investigated how SIS behaves in relation to phoneme learning. We analyzed ERPs in response to Finnish participants' pronunciations on two phonemes (Speak condition): the Estonian phoneme /õ/ (unfamiliar) and the Finnish phoneme /ö/ (familiar). After pronunciation the participants heard an immediate playback of their own vocalizations (Listen condition). We hypothesized that SIS would increase towards the end of the experiment in the Estonian phoneme condition, because the attempted sound and produced sound would match more closely as a result of learning the phoneme. We ran analyses in three time-windows (N1, P2, and Slow-Wave). We assessed learning by having a native Estonian researcher rate the participants' attempts on the Estonian phoneme from 1 (not resembling /õ/ at all) to 4 (excellent pronunciation of /õ/). Based on our behavioral data analysis, our experiment did produce improvements on the Estonian phoneme pronunciations as the trials went on. However, we did not observe any significant changes in ERPs in the N1 time-window or the P2 time-window. These results indicate that the SIS event did not change as the trials moved forward, nor differed between the Finnish and Estonian phoneme conditions. Therefore, phoneme learning did not seem to affect the magnitude of SIS. We found that the ERPs changed as a function of trials in the Slow-Wave time-window for the Estonian phoneme in the Speak condition, turning more positive as trials went on. These results indicate that the brain responds differently to the Estonian phoneme pronunciation compared to the Finnish pronunciation in the Slow-Wave time-window (300-500ms). This effect took place parallel to improvements on the pronunciation, possibly reflecting high-level cognitive processes related to phoneme learning and the production of a new sound.

Key words: Speaking Induced Suppression, SIS, P2, N1, Slow-Wave, EEG, ERP, phoneme learning.

Table of contents

1 Introduction

Speech production is a complex process that involves different brain areas, integrating auditory perception and motor movement. The integration of auditory perception and motor activity is a central concept in speaking, including the process of learning new phonemes. In phoneme learning, sensory processes are involved in the hearing the central characteristics of the sound, and motor processes are involved in the production of the sound. A person must evaluate how well a sound they produced matches the sound they were trying to produce, and if necessary, adjust their pronunciation. The brain shows suppression in the auditory cortex in response to self-produced sounds in comparison to the same sounds that are passively heard, a process that is referred to as Speaking Induced Suppression (SIS) (Niziolek et al., 2013). This suppression seems to be increased when the produced sound matches the attempted sound closely (Ventura et al., 2009). In our study we investigated the mechanisms behind phoneme learning by focusing on SIS and how it changed as the participants learned to pronounce an unfamiliar phoneme. Previous studies focusing on phoneme learning have mainly used listening tasks, and electroencephalography (EEG) correlates during active phoneme pronunciation have received little attention. No other study has focused on the SIS correlate in phoneme learning during active pronunciation.

Shedding light on the role of SIS in speech production mechanisms could potentially further the understanding of the systems underlying our ability to speak. Understanding how these types of neurophysiological functions reflect phoneme learning could be used to develop therapies for people with speech deficits, learning deficits, or even auditory deficits. The aim of this research study was to characterize how SIS behaves in the process of learning to pronounce an unfamiliar phoneme. We did not expect to see changes in SIS in the familiar phoneme condition, because we assumed that the mismatch between the produced and attempted sound for this phoneme was smaller than for the unfamiliar phoneme. Answering these questions would give insight on what mechanisms SIS reflects and how it relates to learning to produce new sounds.

2 Background

2.1 Auditory Processing of Heard and Spoken Phonemes

Basic auditory perception is a crucial component in language acquisition. Early auditory abilities have an impact on language development in normal infants and individuals with language related disorders (Mueller et al., 2012), and the level of speech perception at 6 months predicts language abilities at 2 years old (Tsao et al., 2004). This supports the idea that phonetic perception contributes to language acquisition. Studies have also shown that low level auditory processes, for example brain stem responses in language-impaired children, contribute to the pathological processes of language disorders (Wible et al., 2005). Individual differences in the perceptual abilities of adults have been linked to language-processing abilities in both native and second languages (Mueller et al., 2012). These findings suggest that basic auditory processing has an important role in the process of learning a language, both in infancy and adulthood.

Brain imaging studies have shed light on the functional structures of the human brain, including the mechanisms behind speech perception. There are several brain areas that contribute to speech processing. The left temporal cortex has been identified as one the crucial areas regarding speech perception. When people are presented with speech or non-speech stimuli, activity occurs bilaterally in the primary auditory cortex (Rinne et al., 1999). The left temporal cortex shows language specific activation when participants are asked to pay attention to the phonetic contents of the stimuli. In the study by Rinne et al., (1999), the researchers used mismatch negativity (MMN) EEG component to measure the response to occasional changes in unattended sound stimuli. The MMN is an EEG component that is elicited when the auditory perceptual system detects a mismatch between an expected stimulus and a stimulus that deviates from that neural representation (Diaz et al., 2008; Näätänen et al., 1997). The MMN response is generated by pre-attentive change-detection process in the auditory cortex bilaterally. In the Rinne et al. study, they recorded electrical activation from the brain to unattended sounds which ranged from non-phonetic to phonetic. The study demonstrated that some phonetic information in the auditory stimulus, even when not attended to and with no semantic relevance, is sufficient to activate the speech systems in the left temporal cortex. This activation emerges at an early, pre-attentive stage of sound analysis, around 100-150ms after stimulus onset.

Speaking is a process that involves both sensory perception and motor movement (Guenther & Vladusich, 2009), and when a person is speaking, auditory feedback is used to adjust vocalizations (Greenlee et al., 2011; Niziolek et al., 2013). Both EEG and magnetoencephalography (MEG) studies have shown a diminished amplitude of auditory evoked responses when the participant produces vocalizations in comparison to passive listening of these same vocalizations (Curio et al., 2000; Greenlee et al., 2011; Heinks-Maldonado et al., 2005; Kudo et al., 2004). This observation reflects SIS. A proper interaction between producing a sound and hearing what was produced is crucial in both acquisition and performance of spoken language (Curio et al., 2000). Disturbances in these interactions have been linked to stuttering, aphasia, and even schizophrenic voice hallucinations, but extensive understanding of the auditory self-monitoring of speaking is still underway (Curio et al., 2000).

2.2 DIVA Model

The DIVA model (Directions into Velocities of Articulators) is a computational model that aims to give a quantitative framework for understanding the roles of different brain regions involved in the speech production processes (Guenther & Vladusich, 2011). The DIVA model has been helpful in interpreting experimental results from human speech systems. Producing speech is a complex process that acquires the cooperation of auditory, somatosensory, and motor areas of the cerebral cortex. This complex motor act involves the coordinated activation of nearly 100 muscles in the respiratory, laryngeal, and oral motor systems (Guenther and Hickok, 2015). Because of this, a large network of different brain regions is utilized. Temporal, parietal, and frontal lobes of the cerebral cortex form a functional unit with sub-cortical structures (basal ganglia, brain stem etc.), which together have been termed the speech motor control system. This speech motor control system is engaged even in the simplest of speech tasks, for example reading single syllables (Guenther & Vladusich, 2011).

According to Guenther and Vladusich (2011), the DIVA model operates in the following way: The production of a speech sound (for example a single phoneme) starts with the activation of neurons associated with that sound in the speech sound map. The activation of these speech sound map neurons leads to motor commands from the primary motor cortex. These motor commands arrive via two control subsystems: the feedforward control system and the feedback control system. The feedforward control system projects directly from the speech sound map to the cerebellum and primary motor cortex, where the articulatory control units are located. The feedback control system is slower and involves indirect projections that pass through the sensory brain areas to the auditory cortex.

Speaking is based on the activation of a motor program, termed the forward model. The feedback model works to correct the work of the forward model (Guenther & Vladusich, 2011). When a person speaks, the feedback model gives information about the possible need to adjust the speech. In the heart of the speech production system is a process that compares how well the produced speech matches the intended speech (Guenther & Vladusich, 2011). SIS is an EEG correlate reflecting this phenomenon. The MMN component has been commonly used to assess related processes and mechanisms, as it is thought to reflect how the brain reacts to unexpected stimuli (Wacongne et al., 2012). However, it does not directly reflect the production of speech sounds. SIS can be used to study the process of actively producing sounds and evaluating how well those sounds match the expectation. The SIS correlate has however received significantly less attention in these mechanisms than the MMN. SIS is an important EEG event in relation to the DIVA model and efference copies (covered in next paragraph), since it is believed to reflect some type of a predictive mechanism (Sato & Shiller, 2018), similarly to the MMN.

2.3 Efference Copy and Corollary Discharge Signals

The brain is good at making predictions about the sensory consequences of well-practiced actions. Efference copy refers to the idea that the motor cortex initiates these predictions by making an internal copy of the predicted outputs. This alerts the sensory cortices about the upcoming feedback and allows the changing of response properties (Niziolek et al., 2013). As a result, brain activity that is directed to the incoming sensation is suppressed (Knolle et al., 2019). Efference copies, which are thought to allow the discrimination between self-produced sounds and the external environment (Eliades et al., 2019; Kudo et al., 2004), seem to be very precise (Heinks-Maldonado et al., 2006). For example, EEG studies have consistently shown that the brain shows suppressed auditory responses to self-produced speech in comparison to the same signal that is passively heard (SIS) (Niziolek et al., 2013). The SIS component is most assessed with “talk-listen” research paradigms, using EEG or MEG. SIS is linked to the efference copy mechanisms (Whitford, 2019), and previous studies have shown that when the self-generated sounds differ from the expected sounds, the auditory cortex response is larger than when the self-generated sound matches the expected sound. When these two matches closely, the auditory response is suppressed (Ventura et al., 2009). It is assumed that the better the match is between the prediction of the sensory feedback and the actual observed feedback, the greater the suppression in the auditory cortex is (Niziolek et al., 2013).

Presumably, when a person learns to pronounce a new phoneme, the faulty pronunciations are corrected with the help of efference copies. It is reasonable to assume that when a person is beginning to learn to pronounce an unfamiliar phoneme, the SIS response is less prominent, because the produced sound does not match the attempted sound. In our study, we examined this process with an Estonian phoneme that was unfamiliar to our native Finnish participants. Based on the assumption that the suppression in the auditory cortex is greater when the internal prediction and the produced sound match closely, we would expect to see the SIS response change as a function of trials. Specifically, we would expect to detect more suppression in the auditory cortex during pronunciation in later trials. This is because we assume that as the trials go on, the participants will learn to pronounce the phoneme better. This would mean that the produced sound matches better with the internal prediction of the pronunciation as well.

Niziolek et al. (2013) conducted a study examining how precisely the brain predicts the sensory consequences of our actions. They used MEG to measure the variability of SIS in repeated productions of the same vowel. The participants produced randomized repetitions of three different vowels, and this task was accompanied with a listen condition, where the participants listened to a playback of their utterances. The researchers found that vowels that deviated from the speaker’s average pronunciation produced decreased SIS, suggesting the pronunciation was less accurately predicted by the speech production system. The auditory cortical responses to non-prototypical speech were less suppressed, similarly to responses to speech errors. It is reasonable to assume that these cortical responses are similar in phoneme learning, where the imperfect pronunciation results in a worse match between the motor commands and produced speech. In the study by Niziolek et al., the auditory responses correlated with later corrective movement, which suggests that the suppression may have functional significance for error correction. Because the motor system showed failure to accurately predict less prototypical speech productions, the researchers theorized that the efferent-driven suppression reflects a sensory goal (what is the attempted pronunciation), instead of a sensory prediction (what sound is produced by *these* specific motor commands).

2.4 Previous Studies on Phoneme Learning

Research has shown that the central auditory system transforms in relation to experience. The auditory system reorganizes throughout the lifespan in line with the auditory input that the individual is presented with (Tremblay, 2007). Studies have found that the physiological representation of sound can be changed through training. These training-related changes can accompany improved perception. In animal research, the physiological changes accompanying training have been linked to several different processes, for example greater number of neurons responding in the sensory area, improved neural synchrony, and to processes where training decorrelates activity between neurons (Tremblay, 2007).

In the initial stages of learning a foreign language, the new language is perceived through native language memory traces that are language specific (Tamminen & Peltola, 2015). These native language memory traces develop in early childhood, and by the age of six months, speech sounds are perceived through the native language system. Peltola et al., (2003) studied the development of foreign memory traces and found that Finnish students of English (at an advanced level) did not show native-like MMN responses for target language categories. The researchers also showed that these Finnish students had smaller responses to their mother tongue in comparison to Finnish monolinguals. They suggested that these findings could reflect incomplete learning of English, and that the two language systems might be intertwined. In any case, both the stage of learning and the linguistic context influence second language perception (Tamminen & Peltola, 2015).

Mueller et al. (2012), studied auditory perception in relation to language learning. They concluded in their study that the ability to extract linguistic rules develops early in infancy and seems to be closely linked to discriminatory abilities and auditory mechanisms. The participants included adults and infants, who listened to frequent standard stimuli, and infrequent pitch deviants and rule deviants. Infants who showed a more mature MMN response for the pitch deviants were the only ones who showed an MMN response to the rule deviants. Similarly, the adults who showed larger MMN effects for pitch processing showed evidence of rule learning.

It has been demonstrated in multiple studies that pre- and post-training neurophysiological responses in listening tasks with standard and deviant sound, change in magnitude as perception improves (Näätänen et al., 1993; Tremblay et al., 1998). The time course of these effects is unclear. Tremblay et al., (1998) trained subjects to identify between two different stimuli that differed in voice onset, to examine the time course of learning on a neurophysiological and behavioral level. The training took place over a period of 10 days. The measure of neurophysiological change was the MMN. The participants showed a variety of time courses for behavioral learning, and they all demonstrated significant changes in at least one of the MMN dimensions (duration, area, and onset latency) by day four. The neurophysiological changes always preceded the behavioral changes, and the MMN changes were observed immediately after the first day of training.

The MMN component has been studied also in relation to learning a new phoneme. Diaz et al. (2008), conducted an EEG study assessing the source of individual differences in learning a second language. They measured ERPs from people who were proficient Spanish-Catalan bilinguals but who differed in their mastery of the phonetic contrast /e-ɛ/, which is part of the Catalan language. They wanted to see if the differences stemmed from domain-general psychoacoustic processes or from differences in specific speech perception processes, and addressed these questions by measuring the MMN. Assuming that the size of MMN reflects the strength of perception, it can be used as a measure of perceived change. Therefore, it can be useful in assessing auditory discrimination accuracy in individuals (Diaz et al., 2008; Näätänen et al., 1993). In their study, Diaz et al. suggested that the individual differences to learn phonetic contrasts is not due to the general psychoacoustic abilities. Instead, the researchers showed differences in the sensitivity of individuals to processing phonetic contrasts, which points to a speech-specific origin of the individual variability in mastering a phoneme in a second language. The participants who had mastered perceiving the Catalan phonetic contrast /e-ɛ/ differed from the participants who were categorized as “poor perceivers” of the same phonetic contrast. The “good perceivers” showed larger MMN responses to phonetic stimuli (both native and non-native) than the “poor perceivers.”

Golestani and Zatorre (2004) conducted an fMRI study on brain activity related to phonetic learning. They scanned ten monolingual English-speaking participants while they performed an identification task of a Hindi dental-retroflex nonnative contrast. The participants were scanned twice, both before and after they received five sessions of training on the contrast. Behavioral measurements showed that the subjects improved in their ability to identify the nonnative contrast. The imaging results showed that the same brain areas were active after a successful learning of the nonnative phonetic contrast that are involved in the processing of native contrasts. Interestingly, they also found that the degree of success in learning the nonnative contrast was accompanied by an increased BOLD signal, especially in the classical frontal speech regions. The effects of learning and neural plastic changes have been shown in training studies using EEG as well. Tamminen et al., (2015) trained Finnish subjects to perceive a distinction in the voicing contrast of fricative sounds, which do not belong in the Finnish phonological system. The results showed native-like memory traces after three days of training, as well as substantial changes in the MMN response.

Näätänen and his group demonstrated the existence of language-dependent memory traces in their study in 1997, by focusing on the MMN component. They showed that these memory traces were activated in the processing of speech, but not when equally complex non-speech acoustic stimuli were processed. They measured the MMN in response to a frequent stimulus (Finnish phoneme prototype /e/) and to an infrequent stimulus. The infrequent stimulus was either a Finnish prototype phoneme /ö/, or a non-prototype, the Estonian phoneme /õ/. They found that the MMN was enhanced when the infrequent deviant stimulus was a prototype (/ö/) when compared to the infrequent non-prototype stimulus /õ/. This was only true for the Finnish subjects, for whom the phoneme /ö/ was familiar, and the Estonian phoneme /õ/ was not. The enhancement of the MMN was language specific, and the Estonian participants showed an enhanced MMN in response to /õ/, and not the Finnish phoneme /ö/. Whole-head magnetic recordings were performed, suggesting that the source for these language specific memory traces was in the auditory cortex on the left hemisphere.

Näätänen et al., 1993, also demonstrated the formation of a memory trace for a complex sound in the human brain, by presenting the subjects with a standard sound (not previously familiar to the participants), and a deviant sound. The deviant sound started to elicit an MMN only later in the experiment, and it was not detected in the beginning. This observation was made only in the condition where the participants were paying attention to the possible differences between the stimuli. This suggests that these adaptive changes do not occur in a passive condition and requires effort.

Alain et al., 2007, conducted an EEG study where they measured ERPs while the participants were presented with two phonetically different vowels. The researchers found that the participants' ability to differentiate between the two vowels improved already within the first hour of practice. According to their source analysis, this gradual improvement was accompanied with the enhancement of an early evoked potential, around 130 milliseconds after voice onset, in the right auditory cortex. Additionally, they detected enhancements in the evoked response in a late time window, around 340 milliseconds. This was located in the right anterior superior temporal gyrus and/or inferior prefrontal cortex. These neurophysiological changes were dependent on the participant’s attention levels and occurred only if the practice was continued. Familiarity with the task structure was not sufficient learning to evoke these changes.

3 Aims

Phoneme learning and its neurophysiological correlates have been examined in many experiments as previously described. However, most of the research in EEG studies has focused on the MMN component. Furthermore, literature on auditory processing during active pronunciation is scarce since most of the previous studies have focused on passive listening paradigms. Our experimental design will address this gap in literature, by focusing on SIS during active pronunciation of an unfamiliar phoneme. If tracking the EEG correlates in phoneme learning proved possible, this could shine light on the different mechanisms at play in language acquisition. This type of knowledge could be useful in the development of better therapies for children with problems in speech development and other speech disorders. Other areas, such as research in the field of auditory deficits, could also potentially gain from the possibility to track learning with neurophysiological measures such as SIS.

The aim of this research study was to examine how SIS behaves as a person learns to pronounce a new phoneme. We tracked SIS in two conditions: 1) pronunciation of a familiar phoneme, and 2) pronunciation of an unfamiliar phoneme that was not part of the participants' native language. Both steps were followed by an immediate playback of the participant's pronunciation. We were interested to see possible changes in the magnitude of SIS. We hypothesized that SIS would increase as the participants learned to pronounce the unfamiliar phoneme better. This is because as people learn to pronounce the phoneme better, the neural prediction (attempted sound) and the produced sound match more closely, which would result in a stronger SIS response. We did not expect to see any change in SIS in the familiar phoneme trials, because we assumed this condition would not involve improving on the pronunciation. If the SIS response would change as the experiment goes on and as the subjects learn to pronounce the unfamiliar phoneme, this would suggest that SIS is related to the mechanisms behind phoneme learning. Additionally, we ran analysis in later time-windows, P2 (230-270ms) and Slow-Wave (350-500ms), since SIS occurs in a relatively early N1 time-window (140–180ms), and previous studies have found effects in later time-windows as well (Alain et al., 2007).

4 Methods

4.1 Participants

Twenty people participated in this study (18 female and 2 male). All the participants were Finnish, with normal hearing and no diagnosed learning disabilities or neurological disorders. All the participants were monolingual, their native language being Finnish. The participants were between 18 and 35 years old. All subjects provided informed consent to participate in the study.

4.2 Stimuli and Procedure

We used EEG to measure the electrical activity in the brain in response to pronouncing an unfamiliar phoneme and a familiar phoneme, and passively listening to a playback of the pronunciation. The participants heard a recording of the Estonian phoneme /õ/, or the Finnish phoneme /ö/, in a random order. The Estonian phoneme is not part of the Finnish phonological system and was unfamiliar to the subjects. After hearing the Cue phoneme (Cue condition), the participants tried to repeat it as well as possible (Speak condition). After repeating the sound, they heard a playback of their attempt on the phoneme (Listen condition). The time between the Cue sound and the cue to start pronouncing was 2.5 seconds. The time between the pronunciation and the playback was approximately 3 seconds with some variation depending on the subject’s pronunciation. The time from the playback sound to the next Cue sound was approximately 4 seconds. This process was repeated 50 times in a block, and the experiment consisted of five blocks (250 repetitions in total during the experiment). There was a short break between each block (between 2-5 minutes depending on the subject’s preference), giving the participants an opportunity to rest and stay alert during all the trials.

The stimuli were recorded by having Dr. Pilleriin Sikka, who is a native Estonian, pronounce the vowels /õ/ and /ö/. The recordings of the two phonemes were approximately the same amplitude, pitch, and duration (500ms). The Cue and playback sounds were played to the participants from a TEAC LS-X8 speaker. The participants’ pronunciations were recorded using GXT 242 Lance microphone, and they were saved in wave file format to the computer.

4.3 Electrophysiological Recording

EEG was recorded with 32 passive electrodes placed according to the 10-10 electrode system (EasyCap GmbH, Herrsching, Germany). Surface electromyograms (EMGs) were measured with two electrodes above and below the lips, and below and to the side of the right eye. Reference electrode was placed on the nose. Ground electrode was placed on the forehead. EEG was recorded with a NeurOne Tesla amplifier using 1.4.1.64 software (Mega Electronics Ltd., Kuopio, Finland). Sampling rate was 500 Hz.

The sound stimuli and participants' speech were recorded using a microphone, and its signal was saved as an EEG recording. We did this to accurately mark the onset times for each stimulus and the onset of the participant's own pronunciation. Figure 1 illustrates the use of the microphone signal for our preprocessing steps.

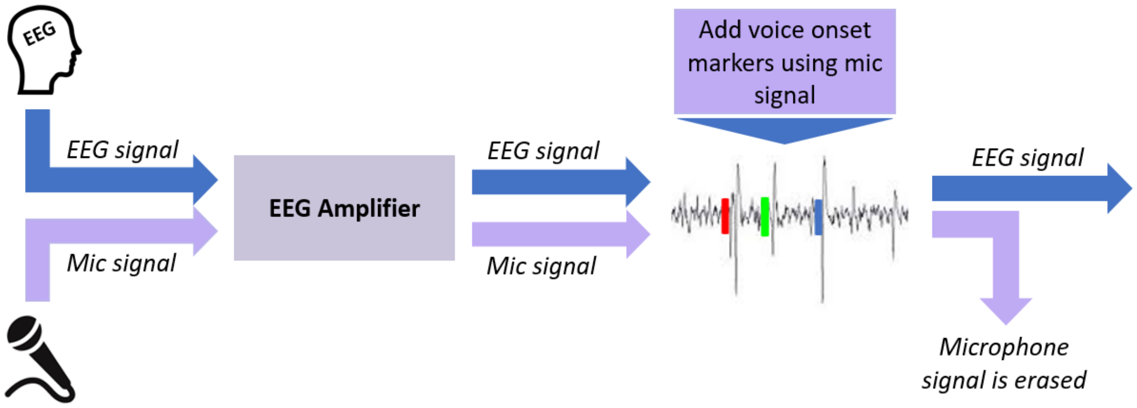
EEG
EEG signal
Mic signal
EEG Amplifier
EEG signal
Mic signal
Add voice onset markers using mic signal
EEG signal
Microphone signal is erased

Figure 1. The microphone was used to record the participants' pronunciations and playback sounds. Microphone signal was saved as an EEG recording and then used to add markers of the voice onset to the EEG signal in each condition (Cue, Speak, Listen). The microphone signal was then deleted.

4.4 Preprocessing

EEG was processed using EEGLAB v14.1.1 software (Delorme and Makeig, 2004) on Matlab 2014b. The sound stimuli were recorded with a microphone that's signal was saved as an EEG recording. First, we high pass filtered the microphone signals recorded with EEG at 100 Hz (to remove noise but keep the sound signal and its transient onset) and used the data to add markers of the stimuli onsets to the continuous EEG data. To determine the onset of stimulus sounds and speech, the signal had to remain above a certain threshold (20 a.u) for ten consecutive samples, and then a marker was added to the time point where the threshold was first crossed. The threshold value was the same for all participants. Markers were added to the time points where the participant heard the /ö/ or /õ/ sound, where they pronounced the sound, and where they listened to the playback of their own pronunciation.

We rejected artifact channels using joint probability of the recorded electrode (EEGLAB pop_rejchan function). The local and global activity probability limit was set at 3 standard deviations. We then interpolated bad channels using pop_interp function, with spherical interpolation, to minimize potential bias in the later average refencing stage. We ran a 1 Hz high pass filter with pop_eegfiltnew function to remove baseline drift, and then removed 50 Hz line noise using CleanLine (bandwidth = 1, winsize = 10, winstep = 10).

For further artifact removal we used Artifact Substance Reconstruction (ASR) method (Chang et al., 2020). We set the cutoff parameter at 20 based on the recommendation by Chang et al. (2020), who concluded in their article that the default values between 5 and 7 removed brain activities too aggressively. We average referenced the data before running Independent Component Analysis (ICA; extended infomax algorithm) to isolate the independent sources underlying the EEG. After ICA we used the DIPFIT plug-in for localizing equivalent dipole locations of the independent components. The rejection threshold was set at 100 (no dipoles were rejected) and two dipoles constrain in symmetry. We used Independent Component Labeling (iclabel function) to add IC classifiers, based on which artefactual components were removed (Pion-Tonachini et al., 2019). Components with residual variance < 15 % and the probability that the component is brain based at > 70 % were considered brain based (i.e. other components were removed).

The data was then split into separate epochs of the different conditions. These conditions included hearing cue /ö/, hearing cue /õ/, speaking phoneme /ö/, speaking phoneme /õ/, listening to playback /ö/, and listening to playback /õ/. The epochs were taken 1 second before the marker, and 1 second after the marker.

Next, we ran a low-pass filter at 40 Hz. Then, we cut the epochs into shorter segments, starting 200 milliseconds before the stimulus onset, and ending 600 milliseconds after stimulus onset. These epochs were used for the statistical analysis. The average number of trials per participant in the Finnish Cue condition was 87 (median = 86, SD = 9.22), and in the Estonian Cue condition the mean was 97 (median = 98, SD = 11.29). The average number of trials in the Finnish Speak condition was 61 (median = 65, SD = 21.66), and in the Estonian Speak condition the average number of trials was 70 (median = 79, SD = 28.25). In the

Finnish Listen condition, the average trial number was 75 (median = 77, SD = 14.55), and in the Estonian Listen condition it was 87 (median = 89, SD = 15.67). The random effects structure in our statistical analysis accounted for the variation between participants.

4.5 Statistical Analysis

We used mixed-effects linear regression analysis to test if Condition (Speak vs. Listen) or Phoneme (Finnish /ö/ vs. Estonian /õ/) factors influenced ERPs at prespecified time-windows and electrodes in single-trial data. In these predictors, the Listen condition and Finnish phoneme were set as baseline categories. In addition, trial number was included in the model as a continuous regressor. Because we were interested in examining if ERP amplitudes changed as the experiment progressed (possibly due to learning), the trial number predictor was not centered, or z scored. The model included all these three predictors and their interactions as fixed-effects models. The model included intercept, condition, phoneme, and condition * phoneme interaction as participant-wise random effects. This means that the regression model considers individual differences in these predictors. The analysis was done on single-trial data, eliminating concerns if individual participants had lower number of trials in some conditions. The analysis was performed for each ERP component (N1, P2, and the Slow-Wave, as described below). We also ran a linear regression analysis to test if phoneme (Finnish vs. Estonian) factor influenced the ERPs at these three time-windows and electrodes in single-trial data, in the Cue sound condition. Finnish phoneme was set as a baseline category and trial number was included in the model as a continuous regressor.

We looked for evidence of learning during the experiment. Learning in this study meant a better pronunciation of the Estonian phoneme /õ/. Learning was assessed by a native Estonian (Dr. Pilleriin Sikka), who listened to the recordings of the participants' attempts on the phoneme. The recordings of each trial were presented to her in random order one participant at a time. The ratings were given on a scale from 1 (not resembling /õ/ at all) to 4 (excellent pronunciation of /õ/). If learning had occurred, we expected to see the trials towards the end of the experiment to be rated higher. For the rating data analysis, we used a fixed-effects linear regression analysis to see if the trial number factor influenced the rating values. We used a random effects structure that accounted for variation between participants when looking at the trial number's effect on the ratings.

We used channels Fz, Cz, FC1, and FC2, on the analysis of N1 and P2 time-windows. The N1 analysis time-window was set between 140–180ms (N1 peak amplitude at 160ms). The P2 analysis time-window was set between 230–270ms (P2 peak amplitude at 250ms). The late time-window analysis, which we termed “Slow-Wave” time-window, was set at 350ms to 500ms. For the Slow-Wave analysis, we included frontal channels F3 and F4, in addition to Fz, FC1, and FC2 channels, because this wave had a more frontal scalp topography (Figure 2). We excluded subject number 11 from the analysis as an outlier, based on the visualization of the ERPs. This participant’s ERPs had large disturbances, showing extremely positive amplitude (500uV) already 1 second before the stimulus onset. The ERPs did not follow any pattern that was observable in the other subjects’ ERP curves.

5 Results

5.1 Description of ERPs

ERPs were calculated starting 200ms before stimulus onset in Finnish and Estonian phoneme conditions. In both Cue and Listen conditions, N1 (negative peak at 160ms) and P2 (positive peak at 250ms) were observed, as shown in Figure 2 and 4. The EEG amplitude was suppressed in the Speak condition, shown in Figure 3. Figure 5 shows the time-course of the auditory stimulus, used to determine the stimulus onset times in ERPs. The figure shows that speech onset was accurately marked on the EEG data in each condition. Our experimental setting successfully produced SIS in the N1 time window, as shown in Figure 6.

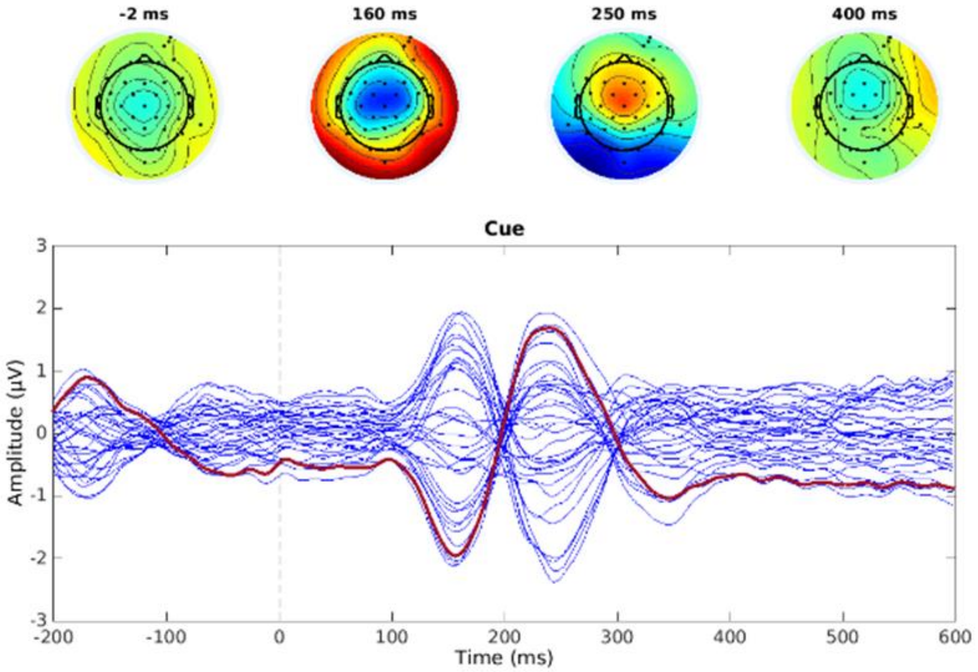

Figure 2. Distribution of electrical activity across the brain in Cue condition. Time-windows: before stimulus onset, at the N1 time window, P2 time window, and Slow-Wave time-window. Below, grand average ERPs from all 34 channels. Mean signal amplitude from central channel cluster (Fz, Cz, FC1, & FC2) highlighted in red.

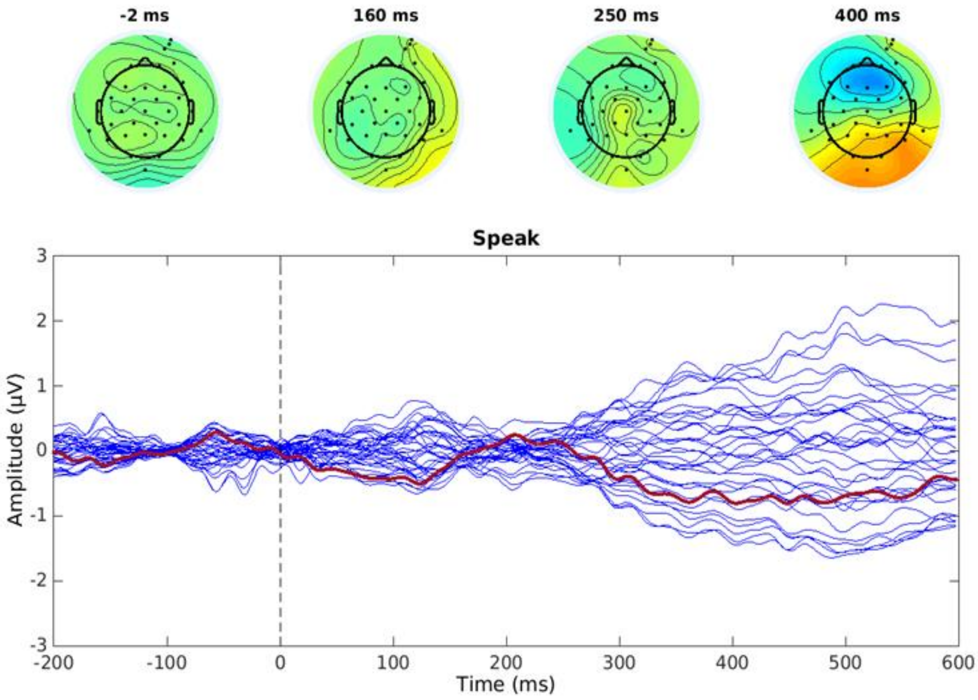

Figure 3. Distribution of electrical activity across the brain in Speak condition. Time windows: before stimulus onset, at the N1 time window, P2 time window, and Slow-Wave time-window. Below, grand average ERPs from all 34 channels. Mean signal amplitude from central channel cluster (Fz, Cz, FC1, & FC2) highlighted in red.

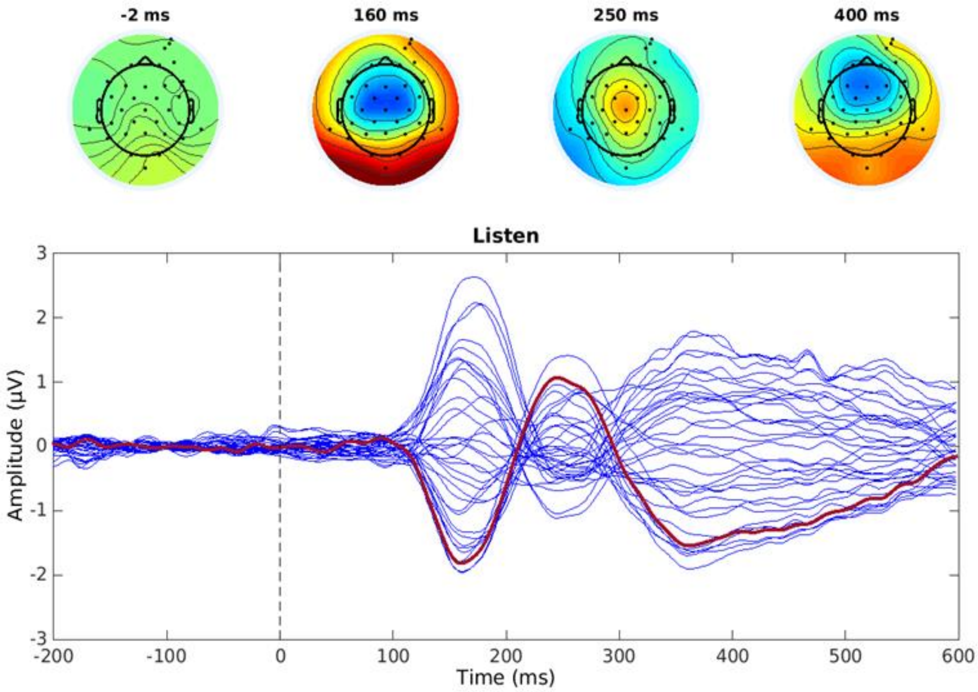

Figure 4. Distribution of electrical activity across the brain in Listen condition. Time-windows: before stimulus onset, at the N1 time window, P2 time window, and Slow-Wave time-window. Below, grand average ERPs from all 34 channels. Mean signal amplitude from central channel cluster (Fz, Cz, FC1, & FC2) highlighted in red.

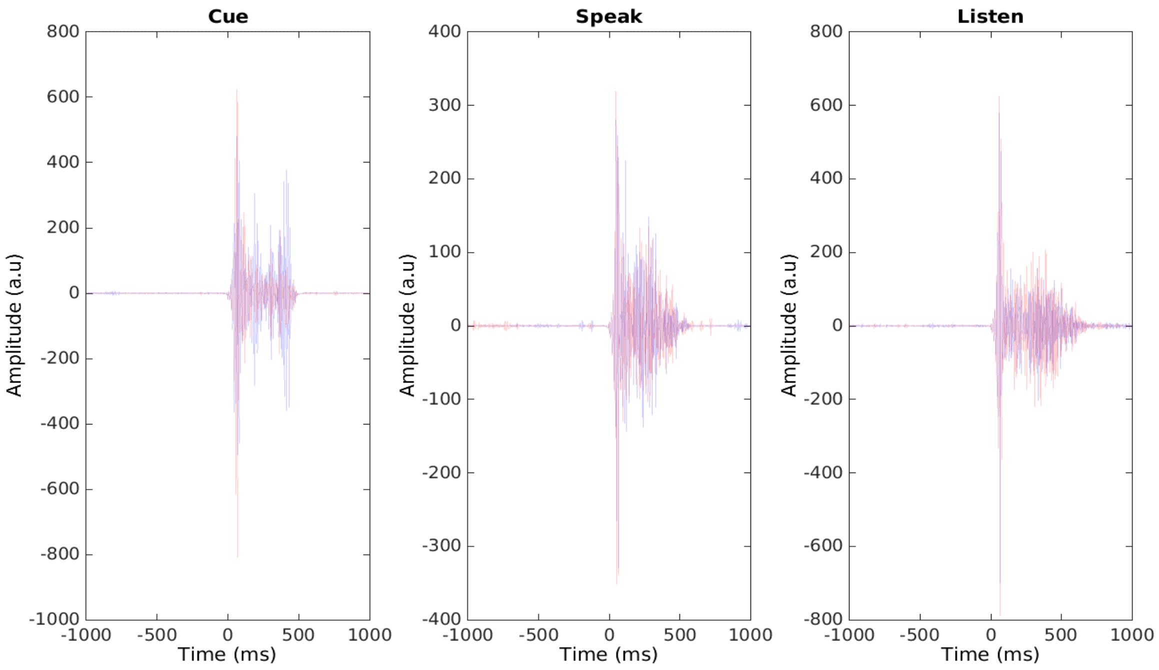

Figure 5. Audio signal measured with EEG in Cue, Speak, and Listen conditions, for Finnish and Estonian phonemes. Graph shows means from all participants. Blue color represents the Finnish trials and red color represents the Estonian trials.

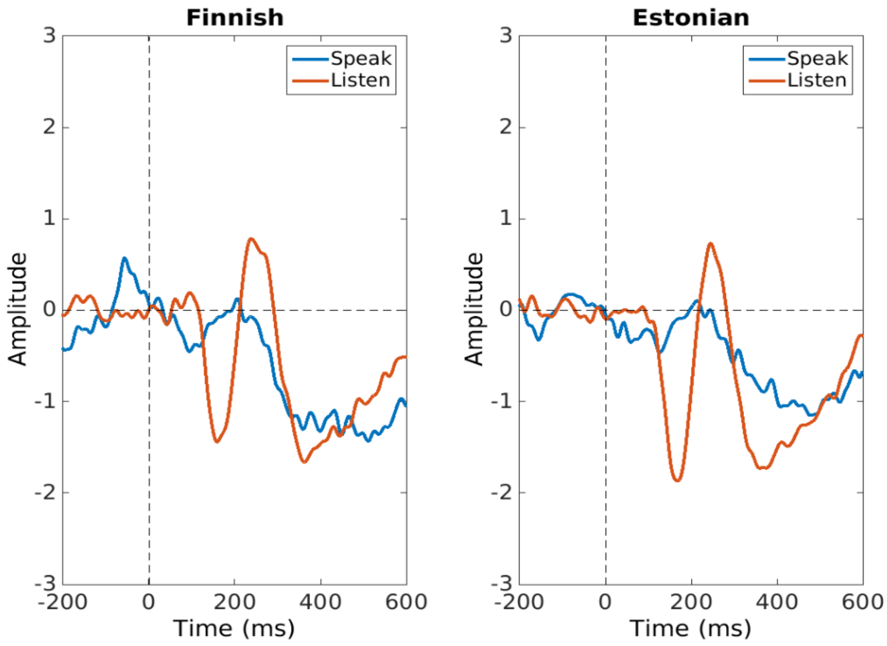

Figure 6. Grand average ERPs from a central electrode cluster (Fz, Cz, FC1, & FC2) in the Finnish and Estonian phoneme conditions. The red and blue lines show the Listen and Speak conditions, respectively.

5.2 Behavioral Data Analysis

Table 1. The results of the fixed effects linear regression analysis on the ratings of each participant's trials on the Estonian phoneme pronunciation.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	1.80	9.66	$1.19*10^{-21}$	1.43	2.16
Trial Number	0.0018	2.30	0.021	0.00027	0.0033

The results of the fixed effects linear regression model on the ratings of each participant's pronunciations on the Estonian phoneme throughout the experiment are shown in Table 1. The ratings ranged from 1 (not resembling /õ/ at all) to 4 (excellent pronunciation of /õ/). The Intercept Estimate value shows the rating at zero trials. The results show that the ratings improved as the trials moved forward (p= 0.021), indicating that learning took place. Figure 7 shows the results of the fixed effects linear regression model, and the individual participants' ratings through the trials.

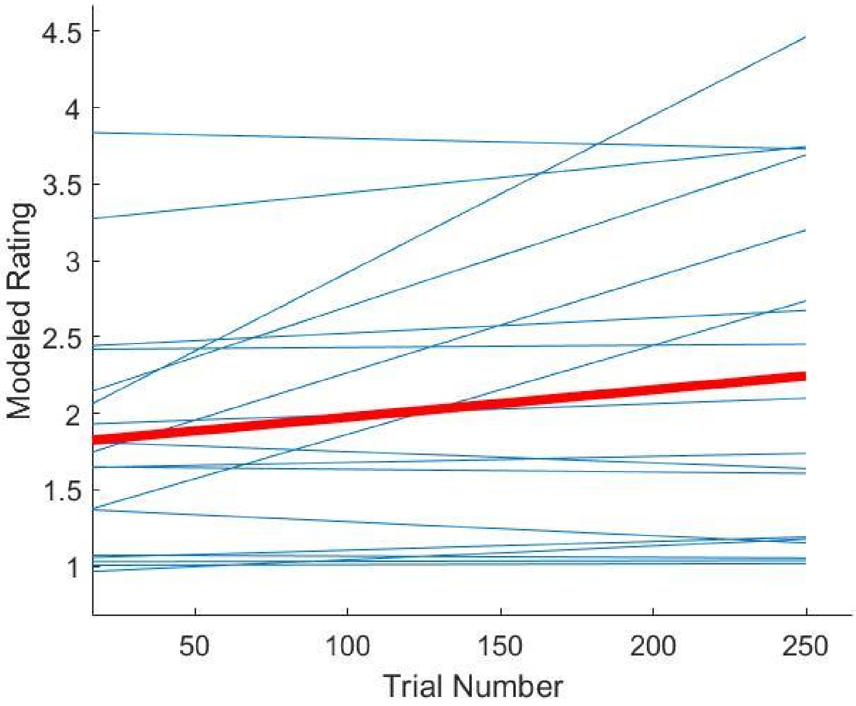

Figure 7. The results of the linear regression model on the ratings on the Estonian phonemes. Blue lines represent each participant's ratings through the experiment. The red line represents the result of the fixed effects linear regression model.

5.3 N1 Time-Window

Table 2. Results of the mixed-effects linear regression model in the N1 time-window.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	-1.52	-5.79	$7.39*10^{-9}$	-2.031	-1.0037
Trials	-0.0015	-0.47	0.64	-0.0079	0.0049
Speak	1.18	3.38	0.00072	0.49	1.86
Estonian	-0.16	-0.602	0.55	-0.66	0.35
Trials: Speak	0.0071	1.36	0.17	-0.00308	0.017
Trials: Estonian	-0.0034	-0.81	0.42	-0.012	0.0049
Speak: Estonian	0.27	0.64	0.52	-0.56	1.089
Trials: Speak: Estonian	-0.0036	-0.54	0.59	-0.017	0.0094

The results of the mixed-effects linear regression model examining ERP amplitudes in the N1 time-window are shown in Table 2. The analysis was done using the central electrode cluster (Fz, Cz, FC1, & FC2). The Intercept indicates the average ERP amplitude in the Listen condition (listening for playback of own voice) for the Finnish phoneme (when trial number equals zero). The Trials predictor indicates the change in amplitude as we move one trial forward. The Speak predictor indicates the change in amplitude from the Listen condition (listening to playback of own pronunciation) to Speak condition (participant pronounced the phoneme) of the Finnish phoneme. As we can see, the Speak condition predicts a statistically significant change in the N1 amplitude towards more positive values, demonstrating SIS ($p = 0.00072$). The Estonian predictor on Table 2 indicates the change in amplitude when listening to the playback of own voice on the Estonian phoneme compared to listening to the Finnish phoneme. As shown in Table 2, listening to the Estonian phoneme did not evoke significantly different ERP amplitudes compared to the Finnish condition ($p = 0.55$). The Trials * Speak predictor shows that the amplitude of Speak condition (Finnish phoneme) did not change significantly as a function of trial number. The Trials * Estonian interaction ($p = 0.42$) was also not statistically significant, indicating that the trials did not predict change in amplitude when the participant was listening to the Estonian phoneme. The Speak * Estonian ($p = 0.52$)

interaction was not statistically significant, indicating that the Speak condition did not affect the amplitude of the ERPs in the Estonian condition. Lastly, the three-way-interaction between Trials * Speak * Estonian was not statistically significant (p = 0.59), indicating that trials did not predict a change in amplitude in the Speak condition compared to the Listen condition for the Estonian phoneme, rejecting our hypothesis that the SIS effect would become more prominent in later trials when pronouncing non-native phonemes.

5.4 P2 Time-Window

Table 3. Results of the mixed-effects linear regression model in the P2 time-window.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	0.94	2.84	0.0046	0.29	1.59
Trials	-0.00095	-0.29	0.77	-0.0075	0.0056
Speak	-1.17	-3.21	0.0013	-1.88	-0.45
Estonian	-0.016	-0.0603	0.95	-0.54	0.51
Trials: Speak	0.0063	1.19	0.23	-0.0041	0.017
Trials: Estonian	-0.0029	-0.68	0.49	-0.011	0.0055
Speak: Estonian	-0.033	-0.076	0.94	-0.89	0.82
Trials: Speak: Estonian	0.0025	0.36	0.72	-0.011	0.016

The results of the mixed-effects linear regression model examining the average ERP amplitudes measured from the central electrode cluster, for the P2 time-window, are shown in Table 3. The difference between Speak and Listen conditions was statistically significant (p = 0.0013), demonstrating suppression in the auditory cortex during the Speak condition, as seen in Figure 6. There were no other significant findings in this time-window. The difference between the Listen condition for the Estonian and Finnish phonemes was not statistically significant (p = 0.95). The amplitude did not change as a function of trials in the Speak condition for the Finnish phoneme (p = 0.23) or for the Estonian phoneme (p = 0.72), rejecting our hypothesis that the brain activity would change in response to pronouncing the Estonian phoneme as the trials went on.

5.5 Slow-Wave Time-Window

Table 4. Results of the mixed-effects linear regression model in the Slow-Wave time-window.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	-1.46	-5.52	$3.44*10^{-8}$	-1.98	-0.94
Trials	0.0024	0.84	0.39	-0.0031	0.0078
Speak	0.5006	1.41	0.16	-0.19	1.19
Estonian	0.022	0.103	0.92	-0.403	0.45
Trials: Speak	-0.0079	-1.78	0.075	-0.017	0.00081
Trials: Estonian	-0.0038	-1.047	0.29	-0.0109	0.0033
Speak: Estonian	-0.26	-0.804	0.42	-0.909	0.38
Trials: Speak: Estonian	0.015	2.604	0.0092	0.0036	0.026

The results of the mixed-effects linear regression model examining the average ERP amplitudes measured from the frontal electrode cluster (F3, F4, Fz, FC1, & FC2), for the Slow-Wave time-window (350ms-500ms) are shown in Table 4. The table shows that the responses to the Finnish phoneme in the Listen condition did not change as a function of trials (p = 0.39) or for the Estonian phoneme (p = 0.29). There was not statistically significant difference in amplitude between the Listen and Speak conditions for the Finnish phoneme in the Slow-Wave time-window (p = 0.16). Activity in the brain was not significantly different when the participants heard their own voice as a playback for the Finnish phoneme compared to hearing the playback of the Estonian phoneme (p = 0.92). The Table shows that the amplitude did not change as a function of trials in the Speak condition for the Finnish phoneme (p = 0.075). However, the Trials * Speak * Estonian Estimate value represents how the amplitude changed in the Speak condition of the Estonian phoneme as a function of trials. This effect is statistically significant (p = 0.0092), suggesting that the ERPs produced by pronouncing the Estonian phoneme changed throughout the experiment in the Slow-Wave time-window, turning more positive. This response is different from the amplitude change in the Speak condition of the Finnish phoneme. The difference between the mean amplitudes in response to the Estonian and Finnish phonemes in the Speak condition in the Slow-Wave time-window, can be seen in Figure 8.

5.6 Cue Analysis: N1 Time-Window

Table 5. Results of the mixed-effects linear regression model in the N1 time-window for the Cue analysis.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	-2.05	-6.84	$9.02*10^{-12}$	-2.64	-1.46
Trials	0.0031	0.0026	0.23	-0.0020	0.0082
Estonian	0.41	0.23	0.078	-0.046	0.87
Trials: Estonian	-0.0035	0.0034	0.30	-0.10	0.0031

Finally, we also examined whether the Finnish and Estonian Cue sounds evoked different ERPs. Table 5 shows the results of the mixed effects linear regression analysis of the average EEG activity across all participants in the Cue condition (hearing the prototype sound of the Finnish phoneme or the Estonian phoneme) in the N1 time-window. The analysis was done using the central electrode cluster (Fz, Cz, FC1, & FC2). As seen in Table 5, the Estonian value is not statistically significant (p = 0.078) indicating that there is no significant difference in the response in the brain for the Finnish and Estonian prototype sounds in the N1 time-window. However, there is a trend of the Estonian Cue sound condition showing a smaller N1 event on average compared to the Finnish Cue sound, although statistically insignificant. Table 5 also shows that the EEG amplitude did not significantly change as a function of trials in response to the Finnish phoneme (p = 0.23) or in response to the Estonian phoneme (p = 0.29), in the N1 time-window.

5.7 Cue Analysis: P2 Time-Window

Table 6. Results of the mixed-effects linear regression model in the P2 time-window for the Cue analysis.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	1.68	4.64	$3.67*10^{-6}$	0.97	2.39
Trials	-0.0024	-0.090	0.97	-0.0055	0.0050
Estonian	0.13	0.51	0.61	-0.38	0.64
Trials: Estonian	-0.0039	-1.11	0.27	-0.011	0.0030

Table 6 shows the results of the mixed effects linear regression analysis of the average EEG activity across all participants in the Cue condition in the P2 time-window. The results show similar trends as in the N1 time-window. The response to the Finnish and Estonian phonemes were not significantly different (p = 0.61). The average EEG amplitude did not significantly change as a function of trials in response to the Finnish Cue sound (p = 0.93) or in response to the Estonian Cue sound (p = 0.27). These results suggest that the Cue sounds did not evoke significantly different responses in the P2 time-window, and that these responses did not change as the trials moved forward.

5.8 Cue Analysis: Slow-Wave Time-Window

Table 7. Results of the mixed-effects linear regression model in the Slow-Wave time-window for the Cue analysis.

Name	Estimate	t value	p value	Lower 95% CI	Upper 95% CI
Intercept	-1.30	-5.87	$4.67*10^{-9}$	-1.73	-0.86
Trials	0.011	4.46	$8.58*10^{-6}$	0.0060	0.015
Estonian	0.44	2.09	0.037	0.027	0.84
Trials: Estonian	-0.0071	-2.27	0.024	-0.013	-0.00093

Table 7 shows the results of the mixed effects linear regression analysis of the average EEG activity across all participants in the Cue condition in the Slow-Wave time-window. This analysis included the frontal electrode cluster (F3, F4, Fz, FC1, & FC2). The results indicate that the Finnish and Estonian Cue sounds evoked significantly different responses (p = 0.037) in the Slow-Wave time-window. The responses to the Finnish Cue sound changed as a function of trials (p = $8.58*10^{-6}$) and responses to the Estonian Cue sound changed as a function of trials (p = 0.024). The regression model shows that the Estonian Cue sound evoked more positive amplitudes than the Finnish Cue sound, but as the trials move forward this difference decreases. After 100 trials this difference has turned to the opposite direction.

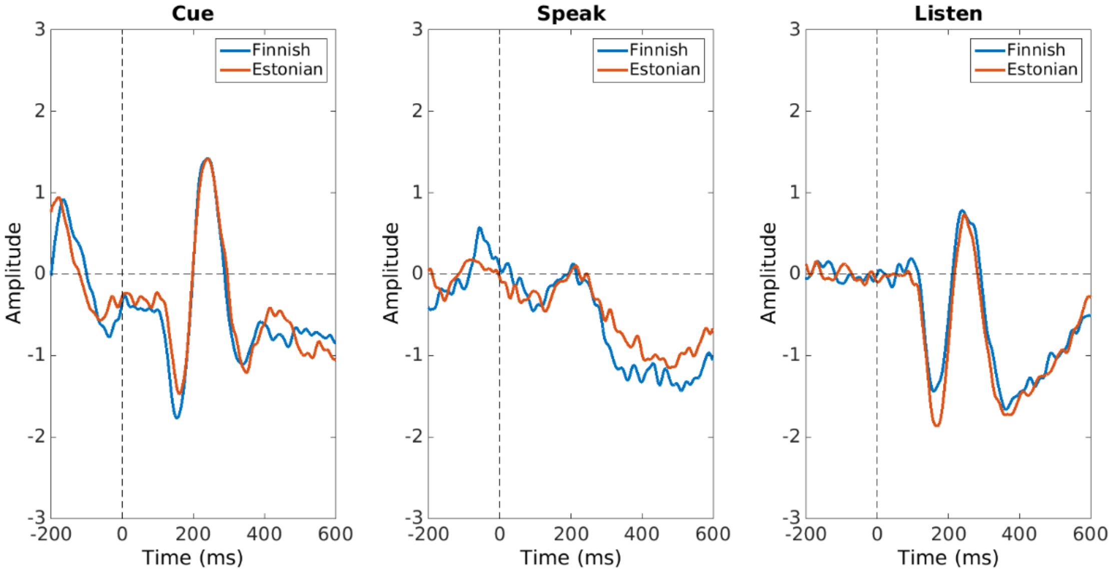

Figure 8. Event-related potentials from a central electrode cluster in the Cue, Speak, and Listen conditions, averaged across participants. The red and blue lines show the Finnish and Estonian conditions.

6 Discussion

In this study we examined how SIS changes as a person learns to pronounce a new phoneme. SIS is thought to reflect a process in the speech production system that compares how well produced speech matches the intended speech (Guenther & Vladusich, 2011), and there seems to be more suppression in the auditory cortex when the produced and attempted sounds match closely (Ventura et al., 2009). We hypothesized that if the participants improved on their pronunciation on the new phoneme, the SIS event would reflect this by growing in magnitude (more suppression in the auditory cortex), and this effect would be different in the Finnish and Estonian phoneme conditions. Our results showed that the ERPs did not differ between the two phoneme conditions and they did not change as a function of trials in the N1 time-window or the P2 time-window, in either Speak or Listen conditions. This result means rejecting our hypothesis that the SIS would change as a function of trials and behave differently for the two phoneme conditions. In the Slow-Wave time-window, we found that the amplitude changed as a function of trials in the Estonian Speak condition, indicating that the response in the brain changed as the trials moved forward while pronouncing the Estonian phoneme. This effect differed between the two phoneme conditions. The amplitude turned more positive for the Estonian phoneme in the Slow-Wave time-window, and this change was not present for the Finnish phoneme. Our behavioral data analysis showed that the participants improved on their pronunciations on the Estonian phoneme as trials moved forward, suggesting that the change in amplitudes in the Estonian Speak condition throughout the experiment could be linked to learning.

6.1 Learning the Estonian Phoneme During the Experiment

Our statistical analysis on the rating data indicated that the participants improved slightly on their pronunciation on the Estonian phoneme as the trials moved forward. We did not observe a significant change in SIS as trials moved forward in the N1 or the P2 time-windows. It is possible that we can only see change in the SIS component when the learning of the phoneme is significantly greater than what the participants achieved during this experiment, even though they did show improvements. The limited sample size can also influence the rating data analysis results, since the participants varied in their learning abilities. However, we did account for this variation with a complex random effects structure in the analysis. With a less strict model, the statistical analysis would have shown an even stronger learning effect. Taking this into account, our statistical analysis suggested that the participants improved in their pronunciations as the trials moved forward, and this observed learning is likely linked to our significant findings.

Previous studies have found that training in specific syllables evoke physiological changes in the MMN response by increasing in amplitude as a person learns to discriminate between syllables (Kraus et al., 1995; Tremblay et al., 1997). It has also been demonstrated that these changes occur quite rapidly. One study showed significant physiological changes in the MMN response after only 45 minutes of training on a syllable discrimination task (Tremblay et al., 1998). This study reported significant differences in the participants' ability to show improvements, and some people required additional training sessions to demonstrate change in the MMN response. Atienza et al., (2002) showed similar results in their study on perceptual learning reflected by the MMN response. There has however been little research focusing on learning in relation to SIS. It is possible that these physiological changes do not occur in the SIS response in the same way they have been observed in the context of MMN.

6.2 N1 Time-Window

Our experiment successfully produced the SIS response, as the amplitude change was predicted by the condition (Speak and Listen). The neuronal activation was significantly suppressed in the Speak condition compared to the Listen condition, showing evidence of a SIS response. Our research question focused on whether SIS would change (if the difference between ERP amplitude in Speak and Listen conditions would grow) as the participants learned to pronounce the Estonian phoneme better. We hypothesized that the Estonian phoneme would evoke a decreased SIS response compared to the familiar Finnish phoneme, and that the SIS effect would change as a function of trials in the Estonian condition. This hypothesis was based on the assumption that SIS reflects a mismatch between the produced sound and the attempted sound. If the participants learned to pronounce the phoneme better, there would gradually be less of a mismatch between the produced and the attempted sound, and this would be observed with a larger SIS response. We did not expect the SIS response to change as a function of trials in the Finnish condition, because this phoneme was familiar to the participants. If SIS is representative of a neural prediction in the brain, it would be expected to remain constant when the participants were familiar with the pronunciation of the phoneme, and the mismatch between produced and attempted sound would have been smaller from the start. SIS did not significantly change as a function of trials in the Finnish condition as expected. However, based on our analysis, we must reject our initial hypothesis in the

Estonian condition. We did not observe a significant change in the SIS response as the trials moved forward for the Estonian phoneme, and SIS did not significantly differ between the Finnish and Estonian conditions.

If SIS does not in fact change in magnitude as a person learns to pronounce a new phoneme, it could be that this event reflects some form of general suppression in the auditory cortex during motor movement, independent from the process of improving on pronunciation. If this is the case, we would not expect to see a difference between phonemes of different familiarity. This would mean that SIS does not reflect a mismatch between the produced and attempted sound. There could however be many factors contributing to the lack of change in the SIS response during our experiment (discussed later), and more research is needed.

6.3 P2 Time-Window

The P2 wave in EEG is the positive deflection peaking around 100-250ms after stimulus onset. Previous studies on auditory feedback in pitch-shifted voice and self-induced sounds have found significant effects both in the N1 time-window and the P2 time-window. Behroozmand et al. (2009) investigated auditory neural responsiveness to self-vocalization, and whether it enhances in response to voice pitch feedback perturbation. They found that the ERP amplitudes in response to feedback perturbation were larger during active vocalization that passive listening, in both P1 and P2 latencies. In a later study Behroozmand et al. (2011) found similar results regarding time-dependent neural processing in auditory feedback in response to self-produced sounds. Based on these previous studies we also ran analyses in the P2 time-window to see if there were any significant effects on the ERPs. In our data the P2 wave was clearly present within the 230-270ms time-window. We found no significant results in this time-window, indicating that the neural responses did not significantly differ between the Finnish and Estonian phonemes for any of the conditions (Speak, Listen, Cue). The amplitudes did not change as a function of trials for Listen or Speak conditions for either Finnish or the Estonian phoneme. Again, we expected to see no change in the ERP amplitudes as a function of trials for the Finnish phoneme, since this was a familiar phoneme which the participants were assumed to have mastered.

If amplitude change in the P2 time-window reflects corrections in pronunciation (Behroomzmand et al., 2011), the lack of change in the ERPs in this time-window, and lack of difference between the two phoneme conditions, could point to the possibility that the participants did not correct their pronunciations. If this is the case, it is not surprising we did not observe significant differences in SIS between the Estonian and Finnish phoneme conditions in the N1 time-window either. The participants simply could have been producing sounds that they meant to produce. Further investigations are needed to better understand what the lack of effects in the N1 and P2 time-windows could mean.

6.4 Slow-Wave Time-Window

We found that there was a significant difference in the ERPs in the brain, when the participants pronounced Finnish versus Estonian phonemes in the Slow-Wave time-window (350ms–500ms). The amplitudes turned more positive as trials went on in the Estonian Speak condition. Alain et al. (2006) found similar results in their study, where they measured ERPs in the brain when the participants were presented with two vowels, and they engaged in a listening task where they tried to differentiate between these vowels. The researchers detected gradual improvements in the participants' ability to differentiate between the two vowels, and this was accompanied with enhancements in the ERPs in the late time-window, around 340ms after voice onset. Importantly, they found that these enhancements were related to the participants' attention levels and occurred only when the practice was continued. This supports our findings that suggest that the change in the ERPs in the Slow-Wave time-window increases as we move forward in trials. We are the first to report these findings for an experiment that focused on ERPs while learning to actively pronounce an unfamiliar phoneme, that we are aware of.

It is likely that the negative Slow-Wave that we observed in this study reflects mechanisms related to phoneme learning, since the ERPs became more positive as the trials moved forward. In past studies, ERPs occurring in later time-windows have been associated with higher cognitive processes, such as phonological processing (Wachinger et al., 2017). Our significant finding parallels our behavioral results which indicated improvements on the Estonian phoneme in later trials. It is also interesting to note that we did not observe similar effects in this time-window for the Finnish phoneme pronunciation, which further supports our theory about learning. The participants were assumed to have mastered the Finnish phoneme pronunciation, and their utterances likely did not change towards the later trials. Since the Finnish pronunciations did not mirror learning, we did not observe changes in the Slow-Wave ERPs either. It is possible these results relate to high-level cognitive processes and learning, where the participants try to fix their pronunciation in the next trial based on the auditory feedback.

6.5 Cue Analysis

The Cue analysis showed that there were no significant differences in the response in the brain for the Finnish and Estonian Cue sounds, in either the N1 or P2 time-windows. This demonstrates that our experiment did not create a significantly different starting position regarding brain activity when the participants started to pronounce either the Finnish or Estonian phoneme in the second stage. The results also show that the ERP amplitudes did not significantly change as a function of trials in response to the Finnish Cue sound or in response to the Estonian Cue sound at the N1 or P2 time-window. This demonstrates that our experimental setting was successful in producing similar responses to the Cue sound in both languages throughout the experiment, and that the Cue sound did not function as a confounding factor. In the Slow-Wave time-window, the Cue stimuli produced an opposite effect on the Estonian * Trials interaction, compared to this interaction in the Slow-Wave time-window in the Speak condition. This suggests that the significant findings in the Slow-Wave time-window in the Speak condition seems to be specific to the self-produced sounds and are likely not explained by the participants changing perception of the Cue sounds.

6.6 Limitations of the Present Study

This study had quite a small sample size of 20 participants, and only two male participants. We set out to recruit between 30 to 40 people, but the COVID-19 pandemic set our study back, leaving us with less participants than planned. This decreases the statistical power of our study. The small sample size can make it harder to detect EEG correlates in phoneme learning, since the participants varied in their ability to learn the new sound. In a small sample size such as this, the results can be easily affected by only a few participants having bad quality data for reasons such as not understanding the assignment, feeling fatigued, or not truly giving effort in the task.

It is possible that the participants did not have enough time to learn to pronounce the Estonian phoneme well. Each participant completed five blocks of repetitions, each block containing 50 trials, lasting 6 minutes. Altogether the experiment lasted 30 minutes for each participant. Our behavioral analysis did indicate that the pronunciation got better during the experiment (although quite slightly), and many of the studies that have reported neurophysiological changes parallel to learning have observed improvements within an hour of training (Alain et al., 2006; Diaz et al., 2008). However, most of the previous studies have focused on learning to differentiate between phonetic contrasts using listening tasks (Alain et al., 2006; Diaz et al., 2008; Mueller et al., 2012; Tamminen et al., 2015). This is quite different from learning to produce a new sound, where the participant must integrate motor learning and auditory perception. This type of learning could take longer than simply differentiating between syllables by listening. Since the participants had only 30 minutes to learn the new phoneme in our experiment, it is possible this was not enough time for improvements that are significant enough to evoke change in the SIS response. However, this does not explain the lack of difference between the Finnish and Estonian conditions and the SIS, because we would still assume that the unfamiliar phoneme would produce a smaller SIS from the start.

The experimental design we used in this study offers artefactual challenges, since the participants are required to move their mouth and jaw while pronouncing the phonemes. Although we did select the Estonian phoneme /õ/ and the Finnish phoneme /ö/ partly because they require very little movement in the jaw and tongue, the motor artifacts could still have an impact on the signal and thus affect our results. However, we employed state-of-the-art preprocessing algorithms to make sure we could clean the motor artifacts from the data the best way possible. Another issue to consider is the possibility of fatigue, which is a common drawback in EEG experiments. If the participants experienced fatigue towards the end of the experiment, it is possible that their ability to improve on the phoneme pronunciation was affected.

Lastly, previous research has shown that individuals differ in their ability to differentiate between syllables in listening tasks, and this has been shown to affect learning and the changes in neurophysiological correlates in previous studies (Diaz et al. 2008). For example, Mueller et al. (2012) found larger mismatch effects in pitch processing for people who showed evidence of rule learning compared to those who did not. It is likely that if people have different abilities in perceiving phonetic contrasts, they also differ in their abilities to learn to pronounce new phonemes. This variability could affect the results of our experiment, especially with a small sample size.

6.7 Further Investigations

Although in this study we did not observe SIS to change as a function of trials, we did find differences in the brain’s electrical activity in response to the different phonemes (Finnish vs. Estonian) that changed as a function of trials in the Slow Wave time-window (350ms-500ms). This result suggests that the brain reacts differently in the process of pronouncing a familiar versus unfamiliar phoneme. This reaction might change as the phoneme becomes more familiar, hence as the trials move forward during the experiment. We cannot make any definitive conclusions based on these results however, since the effect could be the result of some other factor, such as fatigue. It is possible that as the experiment moves forward, the participants experience fatigue faster on the unfamiliar Estonian phoneme, as opposed to the familiar Finnish phoneme. Because of the limitations in this study, future investigation is necessary. Studies could for example look at the effect we found in this study in the context on language development and learning disabilities. It would be interesting to see if the change in the Slow-Wave time-window differs between groups of children where some have issues in language and speech development, and some children are developing normally. If the children suffering from a speech disorder differed in their ERPs in this time-window, for example if their ERPs did not demonstrate the same type of change toward the end of the experiment, we would gain more insight into the possible dysfunctions underlying their condition. This would also provide more evidence that our findings could be related to phoneme learning and higher cognitive functions.

Lastly, subsequent research on this topic should maximize the learning that occurs on the unfamiliar phoneme. Previous studies that have focused on language learning found multiple sessions to be effective in promoting learning (Tremblay, 2007). Furthermore, future investigations should consider that people differ in their abilities to differentiate between phonetics contrasts as well as in their ability to learn (Atienza et al., 2002, Tremblay, 2007). Tremblay et al. (1998) observed in their study that some participants demonstrated improvements after only one or two training sessions while others required additional training sessions before significant perceptual changes became evident. Since people differ in their ability to learn phonemes and their ability to differentiate between phonetic contrasts (Tremblay et al., 1998, Trembley et al., 2007) it is important to increase the participant number in the future to gain more statistical power. This should be taken into considerations in future studies on this topic, and the researchers should account for the possibility of slow learners among the participants.

7 Conclusion

In this study we investigated how SIS changes as a person learns to pronounce a new phoneme. Our behavioral data analysis indicated that participants did learn to pronounce the Estonian phoneme better as the trials moved forward. Although we did not observe any significant changes in SIS in either the N1 or P2 time-windows, we did find that the amplitudes turned more positive as trials went on in the Estonian Speak condition in the Slow-Wave time-window. This result could be attributed to learning to pronounce the Estonian phoneme better, possibly reflecting some form of higher-level cognitive processing while learning to produce a new sound. It is possible that the level of learning in our experiment was not significant enough to induce changes in SIS in the N1 and P2 time-windows, and the small sample size due to the COVID-19 pandemic further decreased our statistical power. Further investigations are necessary to determine what processes the changes in the Slow-Wave time-window reflect and how they relate to learning.

References

Alain, Claude & Snyder, Joel & He, Yu & Reinke, Karen. (2007). Changes in Auditory Cortex Parallel Rapid Perceptual Learning. *Cerebral cortex*, 17(5), 1074-1084.

Atienza M, Cantero JL, Dominguez-Marin E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learn Mem*; 9(3):138–150

Behroozmand, R., Karvelis, L., Liu, H., & Larson, C. R. (2009). Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, *120*(7), 1303–1312.

Behroozmand, R., Liu, H., & Larson, C. R. (2011). Time-dependent neural processing of auditory feedback during voice pitch error detection. *Journal of cognitive neuroscience*, *23*(5), 1205–1217.

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human brain mapping*, *9*(4), 183–191.

Chang, S. -H. Hsu, L. Pion-Tonachini and T. -P. Jung, (2020). Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-Channel EEG Recordings. *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1114-1121.

Díaz, B., Baus, C., Escera, C., Costa, A., & Sebastián-Gallés, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(42), 16083–16088.

Eliades, S. J., & Wang, X. (2019). Corollary Discharge Mechanisms During Vocal Production in Marmoset Monkeys. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, *4*(9), 805–812.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506.

Greenlee, J. D., Jackson, A. W., Chen, F., Larson, C. R., Oya, H., Kawasaki, H., Chen, H., & Howard, M. A., 3rd (2011). Human auditory cortical activation during self-vocalization. *PloS one*, *6*(3), e14744.

Guenther, F. H., & Vladusich, T. (2011). A Neural Theory of Speech Acquisition and Production. *Journal of neurolinguistics*, *25*(5), 408–422.

Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M., & Ford, J. M. (2005). Fine-tuning of auditory cortex during speech production. *Psychophysiology*, *42*(2), 180–190.

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport*, *17*(13), 1375–1379.

Knolle, F., Schwartze, M., Schröger, E., & Kotz, S. A. (2019). Auditory Predictions and Prediction Errors in Response to Self-Initiated Vowels. *Frontiers in neuroscience*, *13*, 1146.

Kraus N, McGee T, Carrell T, King C, Tremblay K, Nicol N. (1995). Central auditory system plasticity associated with speech discrimination training. J Cogn Neurosci, 7:27–32

Kudo, N., Nakagome, K., Kasai, K., Araki, T., Fukuda, M., Kato, N., & Iwanami, A. (2004). Effects of corollary discharge on event-related potentials during selective attention task in healthy men and women. *Neuroscience research*, *48*(1), 59–64.

Mueller J., Friederici A., Männel C. (2012). Auditory perception and language learning. Proceedings of the National Academy of Sciences, 109 (39) 15953-15958.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 33(41), 16110–16116.

Näätänen, R., Lehtokoski, A., Lennes, M. *et al.* (1997) Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature,* 385, 432–434.

Näätänen, R., Paavilainen, P., Tiitinen, H., Jiang, D., & Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology*, *30*(5), 436–450.

Peltola, M. S., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., & Näätänen, R. (2003). Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience Letters*, *352*(1), 25-28.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). The ICLabel dataset of electroencephalographic (EEG) independent component (IC) features. *UC San Diego*.

Rinne, T., Alho, K., Alku, P., Holi, M., Sinkkonen, J., Virtanen, J., Bertrand, O., Näätänen, R. (1999). Analysis of speech sounds is left-hemisphere predominant at 100-150ms after sound onset. *NeuroReport,* 10, 1-5.

Sato, M., & Shiller, D. M. (2018). Auditory prediction during speaking and listening. *Brain and language*, *187*, 92–103.

Tamminen, H., Peltola, M. S., Kujala, T., & Naatanen, R. (2015). Phonetic training and non-native speech perception - New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioral measures. *International Journal of Psychophysiology*, *97*(1), 23-29.

Tremblay K, Kraus N, Carrell TD, McGee T. (1997). Central auditory system plasticity: generalization to novel stimuli following listening training. *J Acoust Soc Am*, 102(6):3762–3773

Tremblay K, Kraus N, McGee T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, 9(16):3557–3560

Tsao FM, Liu HM, Kuhl PK (2004) Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Dev,* 75:1067–1084

Ventura, M. I., Nagarajan, S. S., & Houde, J. F. (2009). Speech target modulates speaking induced suppression in auditory cortex. *BMC neuroscience*, *10*, 58.

Wachinger, C., Volkmer, S., Bublath, K., Bruder, J., Bartling, J., & Schulte-Körne, G. (2017). Does the late positive component reflect successful reading acquisition? A longitudinal ERP study. *NeuroImage. Clinical*, *17*, 232–240.

Wacongne, C., Changeux, J., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience, 32*(11), 3665-3678.

Wible, B., Nicol, T., & Kraus, N. (2005). Correlation between brainstem and cortical auditory processes in normal and language-impaired children. *Brain: a journal of neurology*, *128*(Pt 2), 417–423.