Changes in late frontal event-related potentials to self-produced foreign phonemes correlate with improvements in pronunciation

Henry Railo[1, *], Anni Varjonen[1, 2], Minna Lehtonen[1,5], & Pilleriin Sikka[1, 2, 3, 4]

[1] Department of Psychology and Speech-Language Pathology, University of Turku, Finland
[2] Turku Brain and Mind Centre, University of Turku, Finland
[3] Department of Cognitive Neuroscience and Philosophy, School of Bioscience, University of Skövde, Sweden
[4] Department of Psychology, Stanford University, Stanford, US
[5] Center for Multilingualism in Society across the Lifespan, Department of Linguistics and Scandinavian Studies, University of Oslo, Norway

Abstract

The pronunciation of foreign phonemes is assumed to involve auditory feedback control processes that compare vocalized phonemes to target sounds. The electrophysiological correlate of this process is known as the speaking-induced suppression (SIS) of early auditory evoked activity. To gain insight into the neural processes that mediate the learning of foreign phoneme pronunciation, we recorded event-related potentials (ERP) when participants (N=19) pronounced either native or foreign phonemes. Analyses of single-trial ERPs revealed no differences in SIS between foreign and native phonemes in early time-windows (approx. 85–290 ms). In contrast, the amplitude of the fronto-centrally distributed late slow wave (LSW, 320–440 ms) was modulated by the pronunciation of foreign phonemes. Whereas the self-produced native phonemes evoked a constant amplitude LSW, the LSW evoked by self-vocalized foreign phonemes shifted towards more positive amplitudes across the experiment. Importantly, the LSW amplitude correlated positively with the improved pronunciation of the foreign phoneme. These results suggest that the LSW may reflect higher-order internal monitoring processes that signal successful pronunciation and enable adjustments to future vocalization.

Keywords: Speaking Induced Suppression, event-related potential, ERP, phoneme learning

Introduction

When learning to pronounce novel foreign phonemes, the individual needs to evaluate how well the sound they produced matches the target phoneme, and in case of a mismatch, attempt to improve their phonation. This process is likely based on unconscious speech control processes but may also depend on the individual's conscious evaluation of their phonation. However, relatively little is known about the neural processes that underlie the learning of foreign phoneme production. To elucidate the electrophysiological correlates of learning to pronounce novel phonemes, in this study we examined auditory activity evoked by self-produced and passively heard foreign and native phonemes.

Previous research on the neural processing of foreign phonemes has focused almost exclusively on examining changes in auditory evoked activity elicited by passively heard sounds. Studies employing the oddball paradigm have shown that foreign phonemes elicit weaker mismatch negativity responses than native phonemes (Díaz et al., 2008; Näätänen et al., 1997; Peltola et al., 2003), indicating a poorer ability to discriminate foreign phonemes. Training improves the discrimination of phonemes, and this is reflected in the mismatch response (Tamminen et al., 2015; Tremblay et al., 1998). Phoneme discrimination training

also increases the amplitude of early (N1/P2) auditory event-related potentials (ERPs), which has been interpreted as training-induced neuroplastic changes in the auditory cortex (Alain et al., 2007; Reinke et al., 2003; Saloranta et al., 2020). Studies have also shown that the amplitude of a late (300–500 ms) slow wave to passively heard phonemes correlates with phonemic learning (Alain et al., 2007; Reinke et al., 2003; Saloranta et al., 2020), but the functional role of this correlate is unclear.

Learning to discriminate the acoustic features of passively heard foreign phonemes is obviously important for learning to produce the phoneme. However, the neural processing of passively heard and self-produced sounds differs significantly (Curio et al., 2000; Houde et al., 2002), which is why research should also examine electrophysiological responses to self-produced sounds. Speech production involves an interplay between motor and sensory processing. During speech production, the motor cortex relays "efferent copies" of the feedforward motor commands to the auditory cortex (Hickok et al., 2011; Houde & Nagarajan, 2011; Tourville & Guenther, 2011). This feedback control system allows rapid comparison of how well the produced speech matches the targeted motor commands and enables "online" adjustment of speech. Studies in which speech is artificially perturbed (.e.g, loudness, pitch, or some specific formant) indicate that this mechanism adjusts spoken phonemes in just 100–150 ms (Hain et al., 2000; Houde & Jordan, 1998).

The electrophysiological correlate of this feedback control is the suppression of self-produced auditory evoked activity. Studies show that auditory ERPs in response to self-produced sounds, as compared to when the same sounds are passively heard, are suppressed in the N1 (100–200 ms), and often also in the P2 (200–300 ms), time-windows (Behroozmand et al., 2011; Curio et al., 2000; Heinks-Maldonado et al., 2005; Houde et al., 2002; Railo et al., 2020). This phenomenon is known as the speaking-induced suppression (SIS). Using magnetoencephalography (MEG), Niziolek et al. (2013) observed that the amplitude of SIS (i.e., ERP amplitude difference between self-vocalized and passively heard sounds) tracked variation in phonation: Auditory responses to phonemes that deviated from prototypical phonemes produced a decreased SIS, and the size of SIS predicted corrections to vocalization. Similarly, when speech is artificially perturbed (e.g., by shifting pitch), SIS decreases, suggesting that the brain detects a mismatch between feedforward motor commands and heard speech (Behroozmand et al., 2009; Behroozmand & Larson, 2011; Chang et al., 2013).

The feedback control mechanism may contribute to the process of learning to correctly pronounce foreign phonemes. Pronunciation is adjusted based on the mismatch between the produced and the target phoneme and, during this process, the individual learns to better produce the desired sound. A similar mechanism is assumed to contribute to phonemic learning in children during speech acquisition. After learning the sound of the target speech, the motor commands used to produce the target are fine-tuned based on feedback control (Tourville & Guenther, 2011). However, we are not aware of any studies that have examined whether SIS amplitude tracks the process of learning to produce a novel foreign phoneme.

While SIS is predominantly observed in the N1 and P2 time windows, the comparison of actively spoken and passively heard phonemes sometimes also reveals differences in late time-windows (300–500 ms). Differences in this late time-window are often neglected, likely because they are too late to reflect automatic feedback speech control, but also because there is no theoretical framework within which to interpret these findings. Typically, activity in late time-windows is taken to reflect higher-order conscious attentive processing. While a prominent positive ERP wave peaking between 300 and 500 ms after stimulus onset (P3) is considered a marker of conscious attentive processing (Polich, 2007), in speech tasks, a late

slow wave (LSW) is often observed. Because conscious monitoring (i.e., meta-cognitive evaluation) of self-produced speech is likely to contribute to phonemic learning, and because earlier studies suggest that LSW may correlate with phoneme discrimination learning (Alain et al., 2007; Reinke et al., 2003; Saloranta et al., 2020), it is important to investigate the LSW time-window.

Here, we recorded auditory ERPs when Finnish participants reproduced a native /ö/ phoneme or a foreign Estonian /õ/ phoneme after hearing the target spoken by a native speaker. In a control condition, the participants listened to a recording of the phoneme they had just spoken. Assuming that participants are not able to produce the foreign phoneme as accurately as the native phoneme, we expected to observe a smaller suppression of N1 amplitudes to self-produced foreign phonemes, as compared to passively heard phonemes (i.e., reduced SIS, indicating that the auditory system detects a mismatch between produced and attempted sounds). Furthermore, assuming that participants learn to better pronounce the foreign phoneme during the experiment, we analyzed the trial-by-trial changes in ERP amplitudes. Because previous studies (Alain et al., 2007; Reinke et al., 2003; Saloranta et al., 2020) indicate that not only the N1 time-windows, but also the late time-windows, are modulated by perceptual learning, we also tested for similar effects in the LSW time-window.

Methods

Participants

Twenty-one participants (students at the University of Turku) volunteered for this study. All participants were Finnish, with normal hearing and with no diagnosed learning disabilities or neurological disorders. All participants were monolingual and reported no previous experience in learning Estonian. Two participants were excluded from statistical analyses due to excessive noise in EEG. Thus, the final sample included 19 participants (range 18-35 years; 17 females, 2 males). The study was conducted according to the principles of the Declaration of Helsinki. All participants provided informed consent to participate in the study. The experiment was approved by the Ethics Committee for Human Sciences at the University of Turku.

Stimuli and Procedure

An overview of a single experimental trial is presented in Figure 1A. Participants heard a recording of the Estonian phoneme /õ/, or the Finnish phoneme /ö/, in a random order. We call this the Cue stimulus condition. The phoneme /õ/ is not part of the Finnish phonological system and was therefore unfamiliar to the subjects. After hearing the Cue phoneme, participants attempted to repeat it as well as possible (Speak condition). After repeating the sound, they heard a playback of their produced phoneme (Listen condition). The stimuli were separated by approximately a 2–3 sec. interval. Intertrial interval was about 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 2–5 min break between blocks.

The Cue stimuli were recorded by a native Estonian as the /ö/ and /õ/ phonemes are both part of the Estonian language. The two phonemes were approximately the same amplitude, pitch, and duration (500 ms). The Cue and Listen stimuli were played to the participants from two TEAC LS-X8 speakers placed about 1 meter to left and right of the participant. The

participants' pronunciations were recorded using a GXT 242 Lance microphone, and they were saved in wave file format.

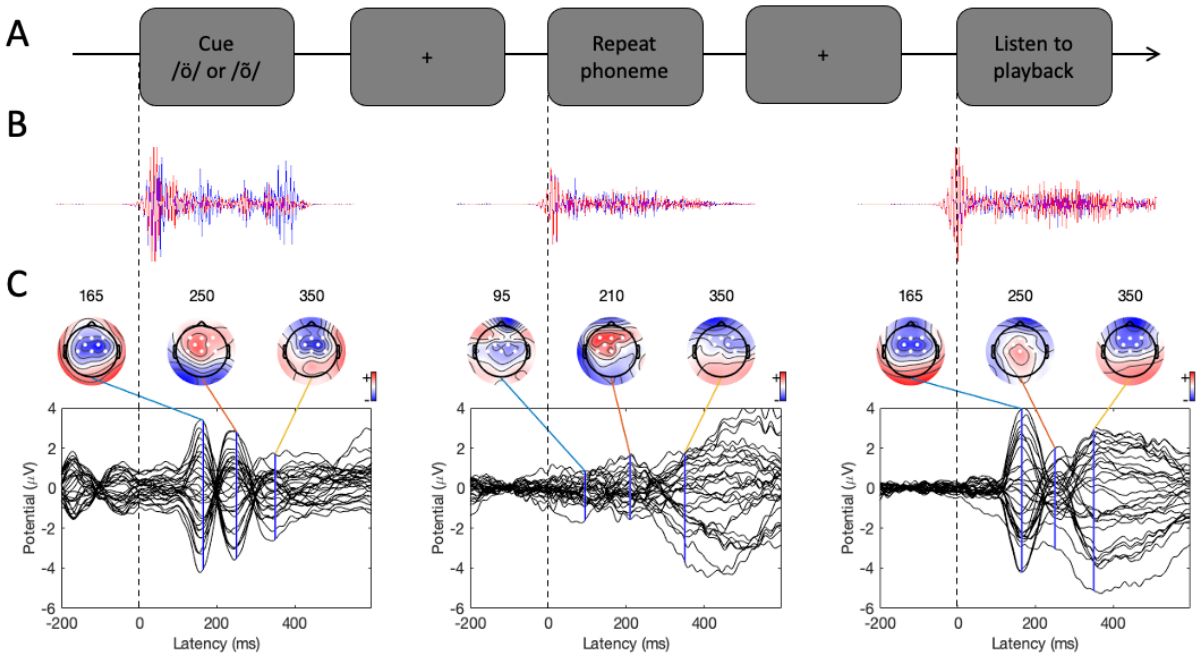

Figure 1. Experimental procedure and ERPs. A) An overview of a single experimental trial. B) Auditory signal recorded with the microphone (average across all participants; blue line = foreign phoneme, red = native phoneme). C) Butterfly ERP plots and scalp maps at three latencies. Scalp color map ranges from -4 (blue) to 4 µVolt (red). White electrode markers indicate the cluster of electrodes on which the statistical analysis was performed.

EEG 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. In addition, the auditory stimuli (Cue, Speak, and Listen conditions) were recorded as EEG signals using a microphone. This allowed us to accurately mark the onset times of auditory stimuli on the EEG.

EEG Preprocessing

EEG was processed using EEGLAB v14.1.1 software (Delorme & Makeig, 2004) in Matlab 2014b. 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 onsets of the stimuli on the continuous EEG data. The onset of auditory stimuli was determined automatically as follows. The microphone signal had to remain above a specified amplitude threshold for ten consecutive samples, and then a marker was added to the sample where the threshold was first crossed. As shown in Figure 1B, this procedure yielded

accurate estimates of phoneme onset times in all three experimental conditions. After this, the microphone channels were removed from the data.

We rejected channels containing artifacts using EEGlab's pop_rejchan function based on kurtosis, spectrum, and probability measures. Then, data were 1 Hz high-pass filtered using pop_eegfiltnew function, and 50 Hz line noise was reduced using the ZapLine plugin (de Cheveigné, 2020). We used Artifact Substance Reconstruction (ASR) to clean continuous EEG using a cutoff parameter at 20 (Chang et al., 2020). We then average-referenced the data and ran Independent Component Analysis (ICA; extended infomax algorithm). After the 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 IClabel to automatically categorize components into brain-based and various non-brain-based categories (Pion-Tonachini et al., 2019). Components with residual variance < 15%, and the probability that the component is brain based > 70%, were considered brain-based (i.e., other components were removed). After this, the removed channels were interpolated.

Next, the data was low-pass filtered at 40 Hz and cut into segments starting 200 milliseconds before stimulus onset and ending 600 milliseconds after stimulus onset. Artifactual trials were removed using the pop_jointprob function (local and global thresholds = 3). The average number of trials per participant in the Finnish Cue condition was 97 (median = 99.5, SD = 9.1) and in the Estonian Cue condition 109 (median = 112, SD = 10.1). The average number of trials in the Finnish Speak condition was 75 (median = 78, SD = 21.0) and in the Estonian Speak condition 88 (median = 99.5, SD = 28.3). In the Finnish Listen condition the average trial number was 85 (median = 87, SD = 14.4) and in the Estonian Listen condition 95 (median = 98.5, SD = 16.4).

Statistical Analyses

We used mixed-effects linear regression analysis to test if Condition (Speak vs. Listen) and Phoneme (Native /ö/ vs. Foreign /õ/) factors influenced ERPs at N1, P2, and LSW time-windows in single-trial data. The analyses were performed on a central-frontal electrode cluster (average of amplitudes across electrodes F3, Fz, F4, FC1, FC2, C3, Cz, and C4, shown in Figure 1C). The analysis was run in Matlab 2014b. The Listen condition and Native phoneme were set as reference categories in the regression models (i.e., intercept is the Native phoneme in the Listen condition). 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). Each condition had its own running trial number (e.g., trial number 10 in Speak/Foreign indicated trial 10 in this specific condition). The trial number regressor was z scored (i.e., intercept of the model represents responses around experiment midpoint). The model included all the three predictors and their interactions as fixed-effects regressors (i.e., Condition × Phoneme × Trial number). The model included the intercept and Speak/Listen conditions as participant-wise random effects because more complex random effect structures (based on Akaike and Bayesian information criteria, AIC/BIC) led to inferior models. Separate models were fitted for each ERP component (N1, P2, and the Slow-Wave, as described below), and the models were pruned by removing outliers.

To examine participants' ability to pronounce the foreign phoneme, recordings of the participants' pronunciation were rated by a native Estonian (author P.S.). During the rating procedure, the recordings were presented 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/). Pronunciation of native phonemes were not rated because all participants could perfectly pronounce them. Because the ratings constitute an ordinal outcome variables, we used mixed-effects cumulative link models to test whether participants' pronunciation changed during the experiment. The analysis was run in R (Version 4.0.2; R Core Team, 2020) using the function *clmm* (logit link model) from the package *ordinal* (Christensen, 2019). The cumulative link model included (single-trial) rating as an outcome variable and trial number as a fixed-effects regressor (trial number was not z scored, i.e., the intercept term represents performance at the beginning of the experiment). Model with by-participant intercept and trial number as random-effects was used because it provided the best-fitting model (based on AIC/BIC).

The data, preprocessing, and analysis scripts are available at https://osf.io/wnd2j/.

Results

Behavioral results

As shown in Figure 2A, on average, participants had difficulties in correctly pronouncing the foreign /õ/ phoneme, although there was also substantial variation between participants. The results of the cumulative link model indicated that, on average, ratings of participants' pronunciation of the foreign phoneme improved as a function of trials (estimate = 0.0049, z = 2.38, p = 0.017). This result is visualized in Figure 2B, which shows the modelled probability of ratings at five different points during the experiment. The lines show the results at six different points during the experiment (black line is the first trial, and the lightest gray line is trial 125 at the end of the experiment).

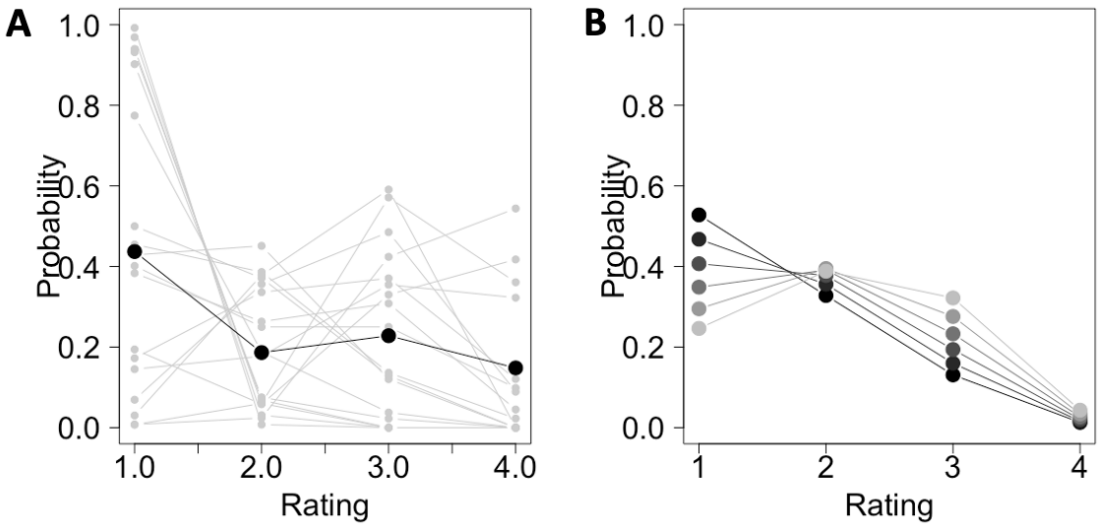

Figure 2. Ratings of foreign phoneme pronunciation. A) Relative frequency of different ratings for the whole sample (black line) and each individual participant (gray lines) across the whole experiment. B) Ratings of participants' pronunciation improved as a function of trial. Lines represent the modelled probability of ratings when trial is 1 (black line), 25, 50, 75, 100, or 125 (lightest gray).

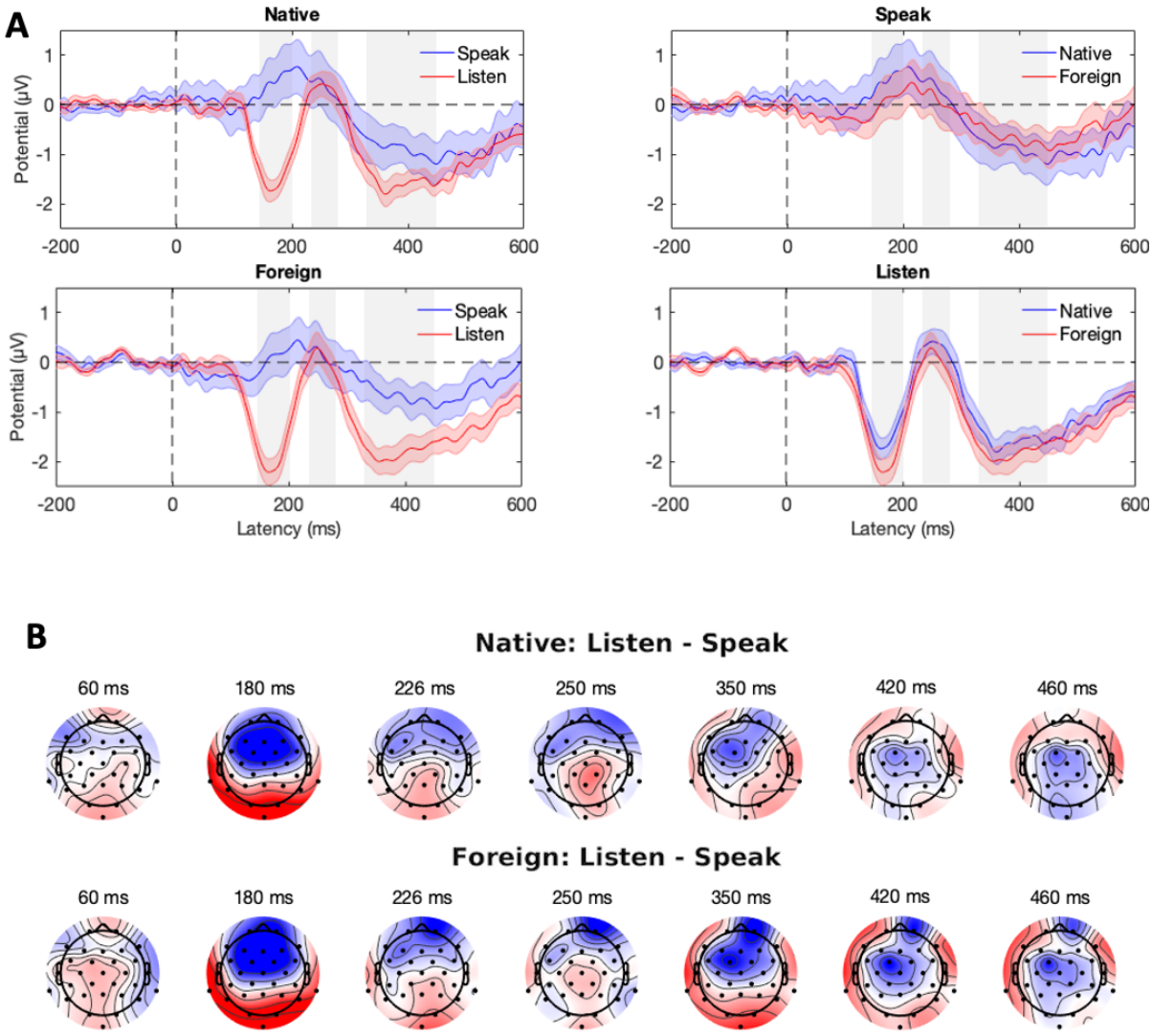

Figure 3. Speaking-induced suppression (SIS), as reflected in the difference of ERP amplitudes between the Listen and Speak conditions. A) Grand-average ERPs on the frontal central electrode cluster (F3, Fz, F4, FC1, FC2, C3, Cz, and C4). The shaded area is the standard error of the mean. The gray rectangles indicate the N1, P2, and LSW time-windows, respectively. B) Scalp maps show the SIS separately for the native and foreign phoneme conditions. Color bar ranges from -2 (blue) to 2 μV (red).

ERP results

Butterfly plots of the ERPs in different experimental conditions are shown in Figure 1C. In both the Cue and Listen conditions, prominent N1 (negative peak at 160 ms) and P2 (positive peak at 250 ms) waves were observed. In addition, the LSW was observed around 350–500 ms after stimulus onset. As expected, the amplitude of the ERPs was suppressed in the Speaking condition relative to the Listen condition, indicating SIS. In addition, the peak latency of the N1 and P2 waves were earlier, and the scalp topography somewhat different in the Speak condition (as compared to the Listen and Cue conditions).

The critical comparison between the Listen and Speak conditions is visualized in Figure 3. Scalp distributions of the difference between the Listen and Speak conditions—reflecting

SIS—are shown in Figure 3B, separately for the Native and Foreign phoneme conditions. We statistically analyzed how the experimental manipulations influenced the N1 (150–200 ms), P2 (230–290 ms), and LSW (320–440 ms) amplitudes. Based on the scalp maps of prominent waves (see Fig. 1C), statistical analyses were performed on the central-frontal electrode cluster (F3, Fz, F4, FC1, FC2, C3, Cz, C4).

N1 amplitudes

The results of the linear mixed-effect regression analyses on N1 amplitudes are presented in Table 1. The intercept represents the average N1 amplitude in the Listen/Native condition. The effect of Trial shows that the amplitude does not change statistically significantly ($p = .17$) as a function of trial. The main effect of the Speak condition indicates that, on average, amplitudes were 1.94 µV more positive in the Speak condition than in the Listen condition ($p < .001$), reflecting the SIS. This effect did not change statistically significantly during the experiment (Trial:Speak interaction, $p = .67$). N1 amplitudes to foreign phonemes were, on average, amplified by 0.38 µV ($p < .001$). The lack of Speak:Foreign interaction ($p = .63$) indicates that the SIS did not differ between the Foreign and Native conditions. Finally, the lack of a three-way Trial:Speak:Foreign interaction indicates that SIS in the Foreign condition was not modulated by trial number ($p = .88$). The results of the model did not change markedly if the model was pruned by removing the (largely redundant) Trial regressor.

Table 1. Results of the mixed-effects regression analysis on N1 amplitudes (df = 6157)

Name	Estimate	SE	t	p	Lower 95%-CI	Upper 95%-CI
Intercept	-1.54	0.19	-7.98	<.001	-1.92	-1.16
Trial	0.12	0.09	1.37	0.17	-0.05	0.29
Speak	1.94	0.45	4.30	< .001	1.06	2.83
Foreign	-0.38	0.12	-3.21	< .001	-0.62	-0.15
Trial:Speak	0.06	0.13	0.43	0.67	-0.20	0.31
Trial:Foreign	-0.19	0.12	-1.62	0.11	-0.43	0.04
Speak:Foreign	0.09	0.18	0.49	0.63	-0.26	0.43
Trial:Speak:Foreign	-0.03	0.18	-0.15	0.88	-0.37	0.32

Because the latency of the N1 wave peaked earlier in the Speak than in the Listen condition, we repeated the analysis with the N1 amplitudes between 84–104 ms in the Speak condition. The overall pattern of results was similar to that reported in Table 1.

P2 amplitudes

The results of the P2 time-window are shown in Table 2. Here, the only marginally statistically significant effect is the reduced P2 amplitude in the Foreign condition ($p = .05$). The results did not markedly change when the model was pruned by removing the Trial regressor.

Table 2. Results of the mixed-effects regression analysis on P2 amplitudes (df = 6143)						
Name	**Estimate**	**SE**	**t**	**p**	**Lower 95%-CI**	**Upper 95%-CI**
Intercept	0.21	0.27	0.79	0.43	-0.32	0.74
Trial	0.01	0.09	0.11	0.92	-0.16	0.18
Speak	-0.09	0.46	-0.20	0.84	-1.00	0.82
Foreign	-0.23	0.12	-1.94	0.05	-0.46	0.00
Trial:Speak	0.20	0.13	1.52	0.13	-0.06	0.45
Trial:Foreign	-0.21	0.12	-1.76	0.08	-0.44	0.02
Speak:Foreign	0.12	0.18	0.67	0.50	-0.23	0.46
Trial:Speak:Foreign	0.11	0.18	0.63	0.53	-0.23	0.45

LSW amplitudes

The results regarding the LSW time-window are shown in Table 3. As indicated by the main effect of Speak condition, the estimated SIS was 0.67 µV in the Native condition (p = .05). In the Foreign condition, the LSW amplitude for passively heard phonemes was slightly (0.26 µV) larger when compared to the Native condition (p = .04). The SIS was 0.43 µV larger in the Foreign condition than in the Native condition (Speak:Foreign, p = 0.02). Finally, the Trial:Speak:Foreign interaction (p = .01) suggests that the difference in SIS between Foreign and Native conditions changed throughout the experiment. The Trial:Speak:Foreign interaction coefficient indicates that, as trial number increased one z unit (roughly 30 trials), the amplitude of the Speak:Foreign interaction increased by 0.49 µV.

Table 3. Results of the mixed-effects regression analysis on LSW amplitudes (df = 6123)						
Name	**Estimate**	**SE**	**t**	**p**	**Lower 95%-CI**	**Upper 95%-CI**
Intercept	-1.83	0.26	-7.01	< .001	-2.34	-1.32
Trial	-0.01	0.09	-0.07	0.95	-0.18	0.17
Speak	0.67	0.34	2.00	0.05	0.01	1.33
Foreign	-0.26	0.12	-2.08	0.04	-0.50	-0.02
Trial:Speak	-0.005	0.13	-0.03	0.97	-0.27	0.26
Trial:Foreign	-0.22	0.12	-1.80	0.07	-0.46	0.02
Speak:Foreign	0.43	0.18	2.35	0.02	0.07	0.78
Trial:Speak:Foreign	0.49	0.18	2.68	0.01	0.13	0.84

Figure 4 visualizes the modelled LWS amplitude and the resulting SIS at three different phases of the experiment. For native phonemes (blue lines), the LSW amplitude, and consequently the SIS, stayed approximately constant throughout the experiment. In contrast, the LSW evoked by the foreign phonemes (orange lines) changed across the trials: whereas in the Listen condition the LSW amplitudes became more negative, in the Speak condition they became more positive. As a result, the SIS increased throughout the experiment (from 0.28 µV to 1.91 µV).

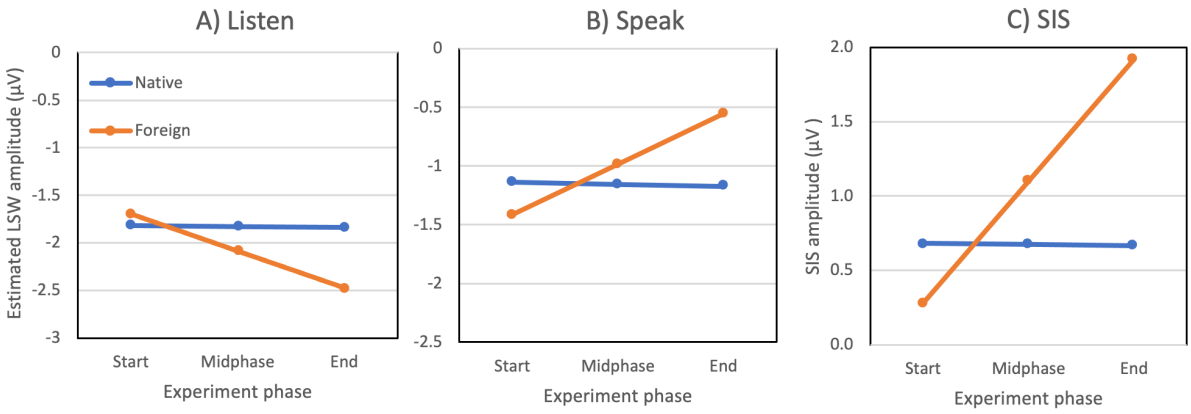

Figure 4. Estimated late slow wave (LSW) amplitudes in the A) Listen and B) Speak conditions, and C) the corresponding speaking-induced suppression (SIS), during different phases of the experiment. The blue line represents the Native phoneme condition and the orange line the Foreign phoneme condition. The SIS evoked by foreign phonemes increased throughout the experiment, whereas it stayed constant for native phonemes. The experiment phase refers to the Trial regressor in Table 3 (i.e., -1.7, 0, and 1.7 z-units at the Start, Midphase and End of the experiment, respectively).

Correlation between pronunciation accuracy and ERPs

Next, we performed mixed-effects regression analyses to examine if differences in N1, P2, and LSW amplitudes were modulated by the pronunciation accuracy (or ratings) of foreign phonemes. The model also included the Speak factor (Speak vs. Listen condition), and the interaction between Speak and pronunciation accuracy (i.e., Rating). The grand-average ERPs are presented in Figure 5A. In the N1 time-window (Fig. 5B), the lack of the main effect of Rating ($p = .26$) and Speak:Rating interaction ($p = .55$) indicated that N1 amplitudes were not modulated by the accuracy of pronunciation. A similar result was obtained when the N1 time-window was calculated based on an earlier time-window (84–140 ms) in the Speak condition. P2 amplitudes (Fig. 5C) did not change as a function of trial in the Listen condition (main effect of Rating, $p = .60$), but P2 was enhanced for phonemes with higher ratings (Speak:Rating, $p = .030$). As shown in Fig. 5D, a similar pattern was observed in the LSW time-window (main effect of Rating: $p = .70$; Speak:Rating interaction: $p = .041$).

Analysis of Cue sounds

We analyzed the ERPs produced by the Cue stimuli using linear mixed-effects regression models. The analysis included a factor indicating Condition (Native vs. Foreign phoneme), running (z scored) trial number, and their interaction. This analysis tests to what extent differences in the Listen and Speak conditions can be observed for stimuli not pronounced by the participants. In the N1 time-window, foreign phonemes produced stronger ERPs ($\beta = -.25$, $t = -1.92$, $p = .054$), and this effect increased through the experiment ($\beta = -.20$, $t = -2.02$, $p = .042$). Models for P2 and LSW amplitudes did not reveal any statistically significant effects.

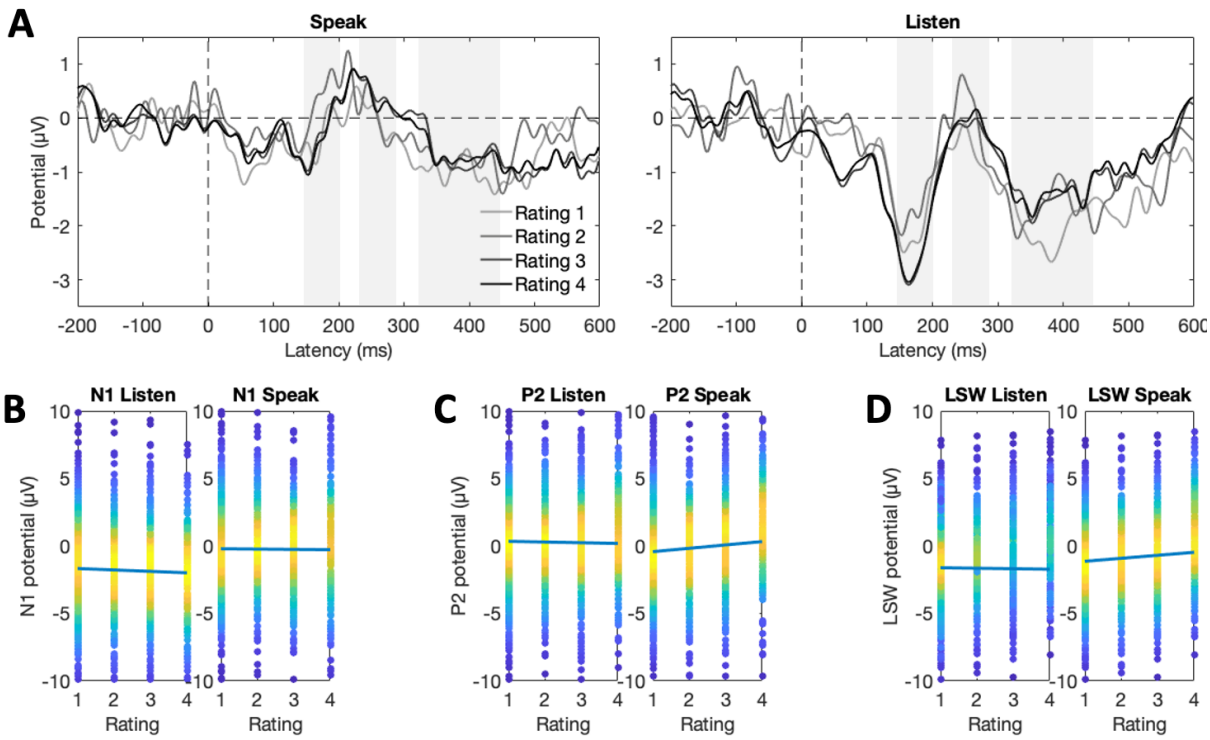

Figure 5. ERPs in the Foreign condition as a function of accuracy of pronunciation (i.e., Rating). A) Grand-average ERPs for different rating categories in the Speak and Listen conditions (frontal-central electrode cluster). The gray rectangles indicate the N1, P2, and LSW time-windows, respectively. B) N1 amplitude as a function of rating, separately for Speak and Foreign conditions. C) P2 amplitude as a function of rating, separately for Speak and Foreign conditions. D) LSW amplitude as a function of rating, separately for Speak and Foreign conditions. In panels B-D the dots indicate single-trial ERP amplitudes, and the color represents the density of the observations (plotted using (Nils, 2021)). The lines in panels B-D display how amplitudes are modulated by rating (results of the linear mixed-effects regression model).

Discussion

When learning to pronounce a novel foreign phoneme, the individual needs to adjust vocalization based on how well their pronunciation matches the target sound. This type of speech monitoring is assumed to depend on auditory feedback control mechanisms (Hickok et al., 2011; Houde & Nagarajan, 2011; Tourville & Guenther, 2011). In the present study, we examined whether a well-known marker of auditory feedback control—speaking-induced suppression (i.e., SIS) of auditory evoked activity—is modulated when participants pronounce foreign vs. native phonemes. We expected to find a reduced SIS for foreign (relative to native) phonemes, indicating a mismatch between produced and target phonemes. In line with a large body of research (Behroozmand et al., 2011; Behroozmand & Larson, 2011; Curio et al., 2000; Heinks-Maldonado et al., 2005; Houde et al., 2002; Knolle et al., 2019; Niziolek et al., 2013), we observed a strong SIS. Whereas we did not find any differences in SIS between foreign and native phonemes in the early or intermediate time-windows (i.e., N1–P2, 85–290 ms after phonation), activity in a later time-window (LSW, starting approximately 300 ms after vocalization onset) was modulated by the pronunciation of foreign phonemes. This effect is particularly interesting because it changed over the course of the experiment. Specifically, the LSW amplitude evoked by self-produced foreign

phonemes became more positive across the trials. Furthermore, the amplitude of LSW to self-produced, but not passively heard, foreign phonemes correlated positively with pronunciation accuracy. In contrast, when participants vocalized native phonemes, the LSW remained constant throughout the experiment. Because the change in the LSW response parallels improvements in pronunciation, and the effect was specific to pronouncing foreign phonemes, this result could offer insight into the neural processes that mediate the learning of a novel phoneme production.

The question arises as to why we did not observe changes in SIS in the N1 and P2 time-windows. Behavioral results indicated that participants had great difficulties in pronouncing the foreign /õ/. This suggests that participants did not yet have a clearly defined acoustic target for the foreign phoneme, and their phonation was likely largely based on the acoustic features of native phonemes. If this is the case, a difference in SIS is not expected because the pronunciation of foreign phonemes relies on similar motor programs as that of native phonemes. However, lack of differences in SIS between foreign and native phonemes should not be taken as evidence that the mismatch between desired and produced speech does not modulate SIS. Many previous studies have shown that, when vocalization does not match the target, SIS is reduced in amplitude (Behroozmand et al., 2009; Behroozmand & Larson, 2011; Chang et al., 2013; Niziolek et al., 2013).

The fact that we did not find any differences in SIS in the early time windows suggests that the participants did not adjust their vocalization of foreign phonemes "online" (i.e., during vocalization). Nevertheless, the participants improved their pronunciation of the foreign phoneme during the experiment, and this change was reflected in the LSW. When the participants listened to the foreign phoneme, the LSW amplitude became more negative across the experiment. This result parallels the findings reported by Alain et al. (2007) and Reinke et al. (2003) who observed that phoneme discrimination training was associated with a modulated LSW (in addition to earlier ERP correlates). But when the participants in the present study actively pronounced the foreign phoneme, the LSW amplitude became more positive across trials. Consequently, the SIS evoked by foreign phonemes increased as a function of trials. Control analyses of the Cue sounds (/ö/ and /õ/ phonemes spoken by a native speaker) indicated that the difference in the LSW amplitudes cannot be attributed to acoustic differences in /ö/ and /õ/ phonemes only.

What cognitive or behavioral processes does the change in the LSW amplitude to self-produced foreign phonemes reflect? First, it could be argued that the change in the LSW amplitude reflects fatigue or inattentiveness, and it applies specifically to foreign phonemes because the pronunciation of native phonemes is easy. However, this explanation is at odds with the finding that the pronunciation of foreign phonemes improved across the experiment. Moreover, if the change in LSW reflects fatigue, LSW should be negatively, not positively, associated with accuracy of pronunciation. Second, it is possible that, just like the SIS in the N1 and P2 time-windows, the LSW evoked by self-produced phonemes is an error signal, indicating a mismatch between the produced and attempted vocalization. However, if LSW reflects an error signal, it should correlate negatively with pronunciation accuracy (Behroozmand et al., 2009; Behroozmand & Larson, 2011; Chang et al., 2013; Niziolek et al., 2013). Instead, in the present study, the enhanced amplitude of LSW indicated improved pronunciation.

SIS can be compared to well-known ERP correlates of performance monitoring (Ullsperger et al., 2014). Correctly performed actions and feedback on a successful performance are associated with a positive amplitude shift in frontal central locations—the reward positivity—

often interpreted as a correlate of reinforcement signals (Carlson et al., 2011; Glazer et al., 2018; Holroyd & Coles, 2002; Hoy et al., 2021; Ullsperger et al., 2014). In addition, the positive amplitude shift in LSW also coincides with the timing and topography of the P3a wave, which reflects attentional capture of salient or motivationally relevant stimuli (Knolle et al., 2019; Polich, 2007). The P3a wave is elicited by sensory stimuli that require a behavioral response form the participants (e.g., participant needs to classify a stimulus and respond using a button press) (Pitts et al., 2012; Scheerer & Jones, 2018), which likely explains why a P3a wave was not observed in the present study. Although we did not observe a P3a wave per se, the neural basis of the positive shift in the LSW amplitude may be similar to the mechanism producing the P3a.

Based on the similarities to reward positivity and P3a, we suggest that the positive shift in the LSW amplitude to self-produced foreign phonemes may reflect saliency or reinforcement signals. These enable one to notice successful pronunciations, which then translates to improved pronunciation during the experiment. By successful pronunciation we do not simply refer to the absence of vocalization errors, but to pronunciation that finds its target *surprisingly* well—so well that it is recruited to drive learning. These successful pronunciations could be "planned" (i.e., in the sense that this was what the individual attempted to do), but they could also be due to pronunciation errors that, by coincidence, match the target sound better than intended. Consistent with this, individuals learn better from feedback following successful trials (rather than errors) (Chiviacowsky & Wulf, 2007). The latter is associated with a positive amplitude shift in frontal electrodes between 200–400 ms that correlates with learning outcomes (Arbel et al., 2013).

The positive shift in the LSW amplitude could also reflect participants' conscious confidence in their performance (assuming that the sources that contribute to P3a also contribute to LSW). Frömer et al. (2021) showed that P3a amplitudes correlated with participants' confidence on the success of their actions. The authors suggest that confidence approximates the amount of noise in the efference copy signals (i.e., their precision), which together with the efference copy and feedback, allow the individuals to make inferences about the success of their action. Their results showed that participants with more accurately "calibrated" confidence (i.e., more accurate estimate of noise in efference copies) learned better (Frömer et al., 2021). This suggests that LSW could reflect higher-order monitoring mechanisms: Whereas SIS in the N1 time-window may reflect automatic "online" corrections based on the efference copy (Hickok et al., 2011; Houde & Nagarajan, 2011; Tourville & Guenther, 2011), the LSW could reflect monitoring processes that integrate multiple sources of information (e.g., perceived auditory feedback on one's pronunciation, predictions based on the efference copy, and the individual's confidence in the accuracy of these predictions). These higher-order monitoring processes could mediate learning by enabling adjustments to *future* vocalizations. Metacognitive evaluation of perception is often associated with the P3 wave, which is assumed to reflect the integration of multiple individual sources of performance related signals (Ullsperger et al., 2010; Wessel, 2012).

In conclusion, our results show that the ability to correctly pronounce a novel foreign phoneme correlates with the amplitude of a late slow ERP (320–440 ms after vocalization). Activity during this time-window was differently modulated when participants pronounced foreign (as compared to native) phonemes, and the effect changed during the course of the experiment, paralleling improvements in pronunciation. We propose that this effect reflects performance monitoring processes that signal successful pronunciations used to adjust future vocalizations. In the future, this correlate could help shed light on various speech-related

phenomena, such as the neural processes underlying native language acquisition and second language learning.

Acknowledgements

M.L. was partly supported by the Research Council of Norway through its Centers of Excellence funding scheme (project number 223265). P.S. was supported by a research grant from the Alfred Kordelin Foundation. We thank Teemu Laine for help with the experimental set up and equipment.

References

Alain, C., Snyder, J. S., He, Y., & Reinke, K. S. (2007). Changes in auditory cortex parallel rapid perceptual learning. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(5), 1074–1084. https://doi.org/10.1093/CERCOR/BHL018

Arbel, Y., Goforth, K., & Donchin, E. (2013). The Good, the Bad, or the Useful? The Examination of the Relationship between the Feedback-related Negativity (FRN) and Long-term Learning Outcomes. *Journal of Cognitive Neuroscience*, *25*(8), 1249–1260. https://doi.org/10.1162/JOCN_A_00385

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*. https://doi.org/10.1016/j.clinph.2009.04.022

Behroozmand, R., & Larson, C. R. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neuroscience*. https://doi.org/10.1186/1471-2202-12-54

Behroozmand, R., Liu, H., & Larson, C. R. (2011). Time-dependent neural processing of auditory feedback during voice pitch error detection. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2010.21447

Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage*, *57*(4), 1608–1616. https://doi.org/10.1016/J.NEUROIMAGE.2011.05.037

Chang, C. Y., Hsu, S. H., Pion-Tonachini, L., & Jung, T. P. (2020). Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-Channel EEG Recordings. *IEEE Transactions on Biomedical Engineering*, *67*(4), 1114–1121. https://doi.org/10.1109/TBME.2019.2930186

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1216827110

Chiviacowsky, S., & Wulf, G. (2007). Feedback after good trials enhances learning. *Research Quarterly for Exercise and Sport*, *78*(2), 40–47. https://doi.org/10.1080/02701367.2007.10599402

Christensen, R. H. B. (2019). *Package "ordinal" Title Regression Models for Ordinal Data*.

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human Brain Mapping*. https://doi.org/10.1002/(SICI)1097-0193(200004)9:4<183::AID-HBM1>3.0.CO;2-Z

de Cheveigné, A. (2020). ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage*, *207*. https://doi.org/10.1016/J.NEUROIMAGE.2019.116356

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Yves Von Cramon, D., & Engel, A. K. (2005). *Behavioral/Systems/Cognitive Trial-by-Trial Coupling of Concurrent Electroencephalogram and Functional Magnetic Resonance Imaging Identifies the Dynamics of Performance Monitoring*. https://doi.org/10.1523/JNEUROSCI.3286-05.2005

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*. https://doi.org/10.1016/j.jneumeth.2003.10.009

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*, *105*(42), 16083–16088. https://doi.org/10.1073/PNAS.0805022105

Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). Response-based outcome predictions and confidence regulate feedback processing and learning. *ELife*, *10*. https://doi.org/10.7554/ELIFE.62825

Glazer, J. E., Kelley, N. J., Pornpattananangkul, N., Mittal, V. A., & Nusslock, R. (2018). Beyond the FRN: Broadening the time-course of EEG and ERP components implicated in reward processing. *International Journal of Psychophysiology*, *132*, 184–202. https://doi.org/10.1016/J.IJPSYCHO.2018.02.002

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research*, *130*(2), 133–141. https://doi.org/10.1007/S002219900237

Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M., & Ford, J. M. (2005). Fine-tuning of auditory cortex during speech production. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2005.00272.x

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/J.NEURON.2011.01.019

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*. https://doi.org/10.1126/science.279.5354.1213

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. In *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/089892902760807140

Hoy, C. W., Steiner, S. C., & Knight, R. T. (2021). Single-trial modeling separates multiple overlapping prediction errors during reward processing in human EEG. *Communications Biology 2021 4:1*, *4*(1), 1–17. https://doi.org/10.1038/s42003-021-02426-1

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. https://doi.org/10.3389/FNINS.2019.01146/BIBTEX

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997).

Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. https://doi.org/10.1038/385432A0

Nils. (2021). *Scatter Plot colored by Kernel Density Estimate* (https://se.mathworks.com/matlabcentral/fileexchange/65728-scatter-plot-colored-by-kernel-density-estimate). Matlab.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? evidence from speech production. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2137-13.2013

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. https://doi.org/10.1016/J.NEULET.2003.08.013

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/J.NEUROIMAGE.2019.05.026

Pitts, M. A., Martínez, A., & Hillyard, S. A. (2012). Visual processing of contour patterns under conditions of inattentional blindness. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_00111

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. In *Clinical Neurophysiology*. https://doi.org/10.1016/j.clinph.2007.04.019

R: a language and environment for statistical computing. (n.d.). Retrieved November 24, 2021, from https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

Railo, H., Nokelainen, N., Savolainen, S., & Kaasinen, V. (2020). Deficits in monitoring self-produced speech in Parkinson's disease. *Clinical Neurophysiology*. https://doi.org/10.1016/j.clinph.2020.05.038

Reinke, K. S., He, Y., Wang, C., & Alain, C. (2003). Perceptual learning modulates sensory evoked response during vowel segregation. *Brain Research. Cognitive Brain Research*, *17*(3), 781–791. https://doi.org/10.1016/S0926-6410(03)00202-7

Saloranta, A., Alku, P., & Peltola, M. S. (2020). Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *147*, 72–82. https://doi.org/10.1016/J.IJPSYCHO.2019.11.005

Scheerer, N. E., & Jones, J. A. (2018). The role of auditory feedback at vocalization onset and mid-utterance. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.02019

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, 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 behavioural measures. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *97*(1), 23–29. https://doi.org/10.1016/J.IJPSYCHO.2015.04.020

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952. https://doi.org/10.1080/01690960903498424

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. https://doi.org/10.1097/00001756-199811160-00003

Ullsperger, M., Fischer, A. G., Nigbur, R., & Endrass, T. (2014). Neural mechanisms and temporal dynamics of performance monitoring. *Trends in Cognitive Sciences*, *18*(5), 259–267. https://doi.org/10.1016/J.TICS.2014.02.009

Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure and Function 2010 214:5*, *214*(5), 629–643. https://doi.org/10.1007/S00429-010-0261-1

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, *14*(4), 593–602. https://doi.org/10.1162/08989290260045837

Wessel, J. R. (2012). Error awareness and the error-related negativity: evaluating the first decade of evidence. *Frontiers in Human Neuroscience*, *6*(APRIL 2012). https://doi.org/10.3389/FNHUM.2012.00088