Speaker-induced Suppression in EEG during a Naturalistic Reading and Listening Task

Garret L. Kurteff, Rosemary A. Lester-Smith, Amanda Martinez, Nicole Currens, Jade Holder, Cassandra Villarreal, Valerie R. Mercado, Christopher Truong, Claire Huber, Paranjaya Pokharel, and Liberty S. Hamilton

Abstract

■ Speaking elicits a suppressed neural response when compared with listening to others' speech, a phenomenon known as speaker-induced suppression (SIS). Previous research has focused on investigating SIS at constrained levels of linguistic representation, such as the individual phoneme and word level. Here, we present scalp EEG data from a dual speech perception and production task where participants read sentences aloud then listened to playback of themselves reading those sentences. Playback was separated into immediate repetition of the previous trial and randomized repetition of a former trial to investigate if forward modeling of responses during passive listening suppresses the neural response. Concurrent EMG was recorded to control for movement artifact during speech production. In line with previous research, ERP analyses at the sentence level demonstrated suppression of early auditory components of the EEG for production compared with perception. To evaluate whether linguistic abstractions (in the form of phonological feature tuning) are suppressed during speech production alongside lower-level acoustic information, we fit linear encoding models that predicted scalp EEG based on phonological features, EMG activity, and task condition. We found that phonological features were encoded similarly between production and perception. However, this similarity was only observed when controlling for movement by using the EMG response as an additional regressor. Our results suggest that SIS operates at a sensory representational level and is dissociated from higher order cognitive and linguistic processing that takes place during speech perception and production. We also detail some important considerations when analyzing EEG during continuous speech production. ■

INTRODUCTION

Background

Speech production and speech perception are frequently studied separately in research, yet the two processes have a robust, interactive theoretical link (Skipper, Devlin, & Lametti, 2017; Houde & Nagarajan, 2011; Tourville & Guenther, 2011; Zheng, Munhall, & Johnsrude, 2010; Watkins, Strafella, & Paus, 2003). Models of the neurobiology of speech production universally include the sensorimotor control of speech, a mechanism by which speakers can detect errors via auditory and somatosensory feedback and subsequently correct those errors (Parrell, Ramanarayanan, Nagarajan, & Houde, 2019; Houde & Chang, 2015; Tourville & Guenther, 2011; Perkell et al., 1997). Errors are identified in part by comparison to the *efference copy*, which represents a set of sensory expectations about the content of an utterance that is generated during pre-articulatory planning of an utterance (Greenlee et al., 2013; Behroozmand & Larson, 2011; Zheng et al., 2010; Hawco, Jones, Ferretti, & Keough, 2009; Hashimoto & Sakai, 2003).

When comparing speech production to perception, a phenomenon known as speaker-induced suppression (SIS) has been observed, where neural responses to (errorless) self-generated sounds are suppressed in relation to externally generated sounds (Brumberg & Pitt, 2019; Niziolek, Nagarajan, & Houde, 2013; Martikainen, Kaneko, & Hari, 2005; Houde, Nagarajan, Sekihara, & Merzenich, 2002). Many EEG (and MEG) studies point to early auditory components such as the N100(m) as a potential biomarker of the efference copy (Behroozmand & Larson, 2011; Heinks-Maldonado et al., 2007; Martikainen et al., 2005). Work in animal models has suggested direct feedback from the motor cortex to inhibitory neurons in the primary auditory cortex suppress responses to self-generated sounds before and during movement (Schneider, Nelson, & Mooney, 2014). One proposed function of SIS is that it is responsible for distinguishing internally and externally generated speech for the purposes of speech motor control (Houde & Nagarajan, 2011; Houde et al., 2002). The degree to which speech is suppressed during SIS is associated with the utterance's adherence to a sensory goal with a greater suppression indicating the acoustic and sensorimotor feedback more closely correspond to the expectations present in the efference copy (Niziolek et al., 2013).

The University of Texas at Austin

https://doi.org/10.1162/jocn_a_02037

Although early evoked potentials (N100/M100) have been identified as biomarkers of the efference copy, the interaction of SIS with other cognitive and linguistic processes ongoing during speech perception and production is not well studied. It is widely accepted that the brain uses some sort of intermediate representations when processing language from its constituent acoustic signal (Mesgarani, Cheung, Johnson, & Chang, 2014; Appelbaum, 1996), and specific representations of the perceptual response may be deemed unnecessary during production (e.g., phonological features, acoustic properties of the speech signal) because of more complete information about auditory stimuli during speech production. Linguistic abstraction of acoustic stimuli into phonological features has been observed during speech production (Cheung, Hamilton, Johnson, & Chang, 2016), but the extent to which feature representations are preserved during SIS has not been directly investigated. In addition, that study mainly addressed changes to feature tuning in the motor cortex itself, rather than changes to tuning in sensory speech areas of the auditory cortex. Importantly, previous research on the content of the efference copy's representations has speculated some form of invariance (i.e., linguistic abstraction). While no explicit source has been identified, several have been posited. Niziolek et al. (2013) identified that the efference copy does not predict all variability in the sensory feedback of speech, suggesting that either the efference copy itself represents an invariant motor plan (cf. the explicit outgoing motor commands of speech), or it represents precise encoding of motor commands that lose their precision (i.e., become invariant) in sensory cortex. Studies of passive listening have established an invariant encoding of noisy acoustic information into linguistic features in auditory cortex (Mesgarani et al., 2014), whereas prior research has established SIS as sensitive to subphonemic variation in acoustic feedback (Niziolek et al., 2013). It is thus unclear whether linguistic encoding is impacted as part of the amplitude reduction in neural response observed during SIS, a process attentive to sublinguistic sensory feedback.

While speech perception does involve feedforward expectations (Poeppel & Monahan, 2011), the forward model of speech perception is not as complete as speech production, because of expectations about utterance content being internally generated during utterance planning. The presence of a forward model during speech production represents a fundamental difference between conditions where speech is suppressed and where speech is not suppressed. Potentially, the feedback monitoring mechanisms at work during SIS could serve a role in general predictive processing; alternatively, the mechanisms of SIS could be specific to speech motor control and any predictability-based modulations could reflect domain-general prediction mechanisms exerting a top–down influence on SIS, which is primarily theorized as a bottom–up sensory process comparing a forward model of motoric goals with auditory/sensorimotor feedback (Niziolek et al., 2013). Studies of altered auditory feedback, in which self-generated speech is acoustically perturbed in real time, show that consistent perturbations of feedback may elicit larger corrective responses than inconsistent ones (Lester-Smith et al., 2020), suggesting that top–down anticipations of feedback may influence how the system responds.

Until recently, EEG studies of speech production were highly constrained in both the content of the produced speech and the analyses available to researchers because of challenges intrinsic to studying speech production that do not hinder the study of speech perception. For example, speech production studies were unable to advance beyond the single word level and frequently were epoched to events other than the onset of articulation (e.g., stimulus presentation, offset of articulation) to prevent EMG artifact associated with articulation from contaminating the neural response (Jiang, Bian, & Tian, 2019; Okada, Matchin, & Hickok, 2018; Singh et al., 2018; Shuster, 2003). Fortunately, advances in artifact correction techniques have resulted in the study of speech production via EEG above the word level. Riès, Pinet, Nozari, and Knight (2021) recently demonstrated the feasibility of analyzing EEG responses to multiword production. Shifting EEG studies toward language as it occurs in natural settings—compared with the heavily constrained single word or syllable-level studies of the past—facilitates generalization to clinical applications and reinforces the interdisciplinary drive to use more ecologically valid stimuli in studies of the neural representation of speech and language (Hamilton & Huth, 2020; Matusz, Dikker, Huth, & Perrodin, 2019). Studies that expand beyond using evoked stimuli and incorporate naturalistic stimuli (e.g., sentences) raise the ecological validity of the research while also providing a window of analysis for the feedforward and feedback processes that link perception and production (Kearney & Guenther, 2019; Houde & Nagarajan, 2011; Poeppel & Monahan, 2011; Casserly & Pisoni, 2010).

Current Study

In this study, we aim to investigate differences in EEG responses between sentence-level speech perception and production, as well as speech perception in consistent and inconsistent contexts to define the neural representations underpinning SIS more precisely. Deviance from a motoric goal has been previously demonstrated to modulate SIS, suggesting the process takes place at a sensory representational level. However, it is unclear whether the suppression of sensory representations during speech production affects linguistic abstractions that are generated from sensory processing. If linguistic representations were suppressed in conjunction with SIS, we would expect to see differential phonological tuning to specific speech features (Desai et al., 2021; Khalighinejad, da Silva, & Mesgarani, 2017; Di Liberto, O'Sullivan, & Lalor, 2015). In addition, we opted to structure our task such that participants can anticipate when playback of auditory responses is inconsistent with the preceding speaking trial. This allowed us to determine if there is a link between predictive processing in passive listening and the consistency or inconsistency of feedback (Lester-Smith et al., 2020). To investigate these questions, we designed an experiment that used identical acoustic stimuli in separate speech perception and production conditions then compared the difference in ERPs as well as in tuning of phonological features across conditions. We hypothesized that, although speech production will be suppressed relative to perception in this study, phonological feature tuning would remain stable between modalities of speech. In addition, we expect a similar trend in inconsistent perceptual stimuli, such that phonological feature representations will remain stable but reduced in amplitude in comparison to consistent perceptual stimuli. If responses to consistent versus inconsistent playback show a similar pattern of suppression to that observed during speaking versus listening, we may conclude that these processes involve similar underlying computations. On the other hand, if we see differences in these patterns, this suggests that SIS is computationally distinct from other phenomena that involve forward modeling or expectations of upcoming speech. In addition, observing differences in linguistic abstraction of acoustics into phonological features can help contextualize the phenomenon in relation to other cognitive and linguistic processes operating during speech perception and production.

METHODS

Participants

Twenty-one participants (11 women, age 24.4 ± 3.9) were recruited from The University of Texas at Austin. This is in line with sample sizes of recent EEG studies of speech production (Riès et al., 2021; Goregliad Fjaellingsdal et al., 2020; Zhao & Rudzicz, 2015). All participants were native speakers of English with typical hearing as assessed through pure-tone audiometry and a speech-in-noise hearing test (QuickSIN, Interacoustics). Participants provided written consent for participation in the study and were compensated at a rate of $15/hr with an average session length of 2 hr (1 hr for setup, 1 hr for recording EEG). One participant was excluded because of a recording error, leaving 20 participants in the final analysis. All experimental procedures were approved by the institutional review board at The University of Texas at Austin.

Materials

The task was designed using a dual perception-production block paradigm, where trials consisted of a dyad of sentence production followed by sentence perception. In each trial, participants overtly read a sentence and then listened to a recording of themselves reading the produced sentence. Perception trials were divided into blocks of consistent and inconsistent stimuli. *Consistent* stimuli consisted of immediate playback of the production trial, whereas *inconsistent* stimuli consisted of a randomly selected production trial from the previous block. Consistent and inconsistent playback trials were presented in a block paradigm to avoid eliciting an "oddball" response, a commonly observed ERP component that elicits a response to randomly deviant perceptual stimuli (Barry, Kirkaikul, & Hodder, 2000). A schematic is provided in Figure 1A. The generation of perception trials from the production aspect of the task allowed stimulus acoustics to be functionally identical across conditions. Sentences were taken from the MultiCHannel Articulatory database, a corpus of 460 sentences that include a wide distribution of phonemes and phonological processes typically found in spoken English (Wrench, 1999). These sentences have been used previously in intracranial studies of speech production (Chartier, Anumanchipalli, Johnson, & Chang, 2018). A subset of 50 sentences (100 for the first two participants) from MultiCHannel Articulatory were chosen at random for the stimuli in the present study; however, before random selection, 61 sentences were manually removed by an author (G. L. K.) for either containing offensive semantic content or being difficult for an average reader to produce to reduce extraneous cognitive effects and error production, respectively. We changed the sentence set from 100 to 50 sentences after the first two participants because of concerns about participant fatigue during the task. Participants completed six blocks of the task for 300 perception and 300 production trials per participant (400 for the first two participants). Sentences had a median length of 2.9 sec. A broadband click tone was played in between trials as an additional cue to assess the effect of EMG correction on low-level auditory responses (see Appendix).

Stimuli were presented in a dimly lit sound-attenuated booth on an Apple iPad Air 2 using custom interactive software developed in Swift (Apple XCode Version 9.4.1). Auditory stimuli were presented at a comfortable listening level via foam-tipped insert earbuds (3 M, E-A-Rtone Gold 10 Ω). Visual stimuli were presented in a white font on a black background after a 1000 msec fixation cross to minimize visual artifact in the EEG signal (Figure 1D, 1E). Accurate stimulus presentation timing was controlled by synchronizing events to the refresh rate of the screen. The iPad was placed on a table over the participants' lap so they could advance trials during the task with minimal arm movement. Participants were instructed to complete the task at a comfortable pace and were familiarized with the task before recording began. Trial information, including onset and offset of each trial, transcriptions of produced and heard sentences, trial type, trial number, and block number were collected by an automatically generated log file to assist in data processing.

EEG Data Collection

Sixty-four-channel scalp EEG and audio were recorded continuously via BrainVision actiChamp amplifier (Brain

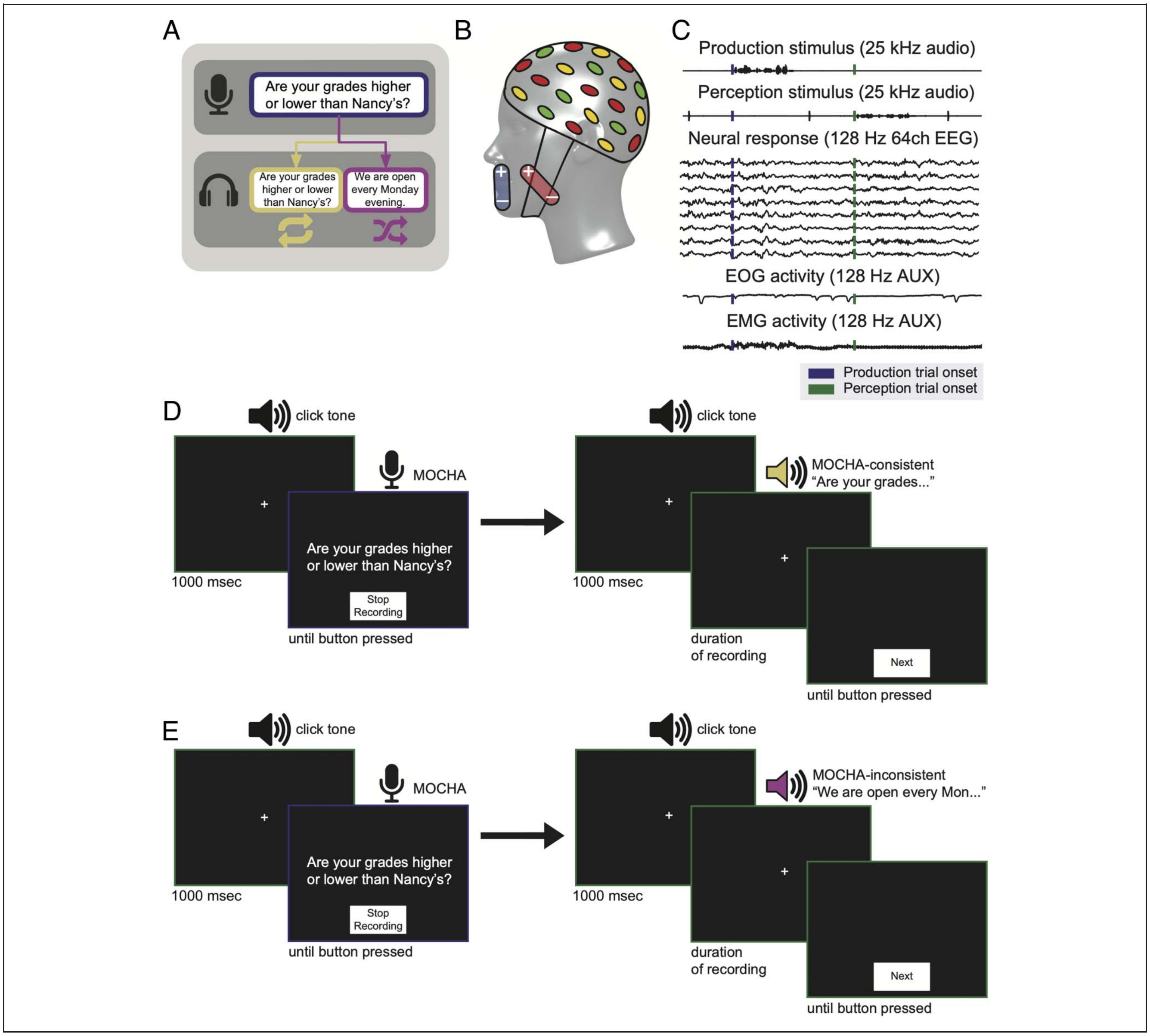
A
Are your grades higher or lower than Nancy's?
Are your grades higher or lower than Nancy's?
We are open every Monday evening.
B
+
−
+
−
C
Production stimulus (25 kHz audio)
Perception stimulus (25 kHz audio)
Neural response (128 Hz 64ch EEG)
EOG activity (128 Hz AUX)
EMG activity (128 Hz AUX)
Production trial onset
Perception trial onset
D
click tone
+
1000 msec
MOCHA
Are your grades higher or lower than Nancy's?
Stop Recording
until button pressed
click tone
+
1000 msec
MOCHA-consistent
"Are your grades..."
+
duration of recording
Next
until button pressed
E
click tone
+
1000 msec
MOCHA
Are your grades higher or lower than Nancy's?
Stop Recording
until button pressed
click tone
+
1000 msec
MOCHA-inconsistent
"We are open every Mon..."
+
duration of recording
Next
until button pressed

Figure 1. Dual perception-production task and EEG data collection schematic. (A) Schematic of trial types in the task. The participant first reads a sentence aloud (indigo) then hears playback of the same audio (yellow, consistent playback condition) or audio from a different random trial (magenta, inconsistent playback). (B) Schematic of auxiliary EMG electrode placement on orbicularis oris (blue) and masseter (red). (C) Visualization of all signals recorded during task, including produced audio (speech), perceived audio (clicks and speech), and EOG and EMG channels. Only eight EEG channels are visualized here, but 64 were recorded and used in analysis. Vertical lines denote the onset of a production (indigo) or perception (green) trial (i.e., the acoustic onset of the first phoneme of the sentence). Blinks are observed as deflections in the EOG channel; muscle activation during production is notable as high activity in the EMG channel. (D, E) Outline of trial procedure for consistent (yellow) and inconsistent (magenta) blocks.

Products) with active electrodes at 25 kHz. A high sampling rate was used to synchronize task audio and EEG, which were recorded using the same amplifier. Conductive gel (SuperVisc, EASYCAP) was applied to the scalp at each electrode, and impedance at each electrode was kept below 15 kΩ throughout the recording. Audio signals from both the insert earphones (presented audio) and microphone (produced audio) were captured as additional EEG auxiliary channels and were aligned with neural data via a StimTrak processor (Brain Products). VEOG was captured via auxiliary electrodes above and below the left eye in line with the pupil. Auxiliary electrodes were also used to capture facial EMG activity (Figure 1B); these electrodes were placed on the orbicularis oris and mandible in the majority of participants (n = 11), but on other muscles important to articulation (masseter [n = 6], submental triangle [n = 2]) in several participants (Stepp, 2012; Van Eijden, Blanksma, & Brugman, 1993; Rastatter & De Jarnette, 1984). Multiple placements were utilized because of issues with electrode adherence caused by participant facial hair. All placements were trialed on a participant who consented to additional time during setup. A reference electrode for all auxiliary electrodes was placed on the left earlobe. Auxiliary EMG placement was not required for preprocessing but provided validation that EMG artifact was removed during preprocessing. EMG activity associated with the onset of articulation, which caused the largest artifacts in the temporal window of interest for ERP analysis of speech production, was automatically detected and epoched from auxiliary EMG channel activity. All recorded signals timed according to stimulus onsets are visualized in Figure 1C. The first two participants did not have auxiliary electrode placement because of unavailability of recording hardware, so EMG activity was corrected based on EEG channels only (see below).

EEG Data Processing

All EEG processing was performed offline using custom Python scripts and functions from the MNE-python software package (Gramfort et al., 2014). EEG, EOG, and EMG data were downsampled from 25 kHz to 128 Hz before analysis. EEG data were referenced to the linked mastoid electrodes (the average of the TP9 and TP10 channels) and notch filtered at 60 Hz to remove line noise. For one participant (OP17), one reference electrode was a bad channel and was interpolated before rereferencing. The data were next filtered from 1–30 Hz (Hamming window, 0.0194 passband ripple with 53-dB stopband attenuation, 6 dB/octave). Bad channels and segments were manually rejected, then Independent Component Analysis (ICA) was performed to correct for EOG and electrocardiographic artifact with the number of components equal to the number of good channels. ICA components related to VEOG, HEOG, and electrocardiographic artifact were manually identified and removed via scalp topography and epoching component activity to vEOG activity (obtained via MNE function create_eog_epochs). The selected ICA components were next removed from the unfiltered data. After ICA, data were filtered at 0.16 Hz and corrected for EMG artifact via blind source separation algorithm based on canonical correlation analysis (CCA; De Clercq, Vergult, Vanrumste, Van Paesschen, & Van Huffel, 2006), a technique that has been previously demonstrated to correct for EMG artifact in speech production EEG tasks (Riès et al., 2021; Riès, Janssen, Burle, & Alario, 2013; de Vos et al., 2010). In line with these studies, CCA was performed in two passes: first, a 30-sec window to remove tonic muscle activity; second, a 2-sec window to remove rapid bursts of EMG associated with speech production. CCA was performed using the Automatic Artifact Removal plugin for EEGLab (Gómez-Herrero, 2007). Validity of CCA artifact correction for the removal of EMG from continuous speech production data is not discussed further in this article; however, an additional verification of the technique can be found in the Appendix. After CCA and before analysis, bad channels were interpolated and data were bandpass filtered between 1 and 30 Hz.

ERP Analysis

Accurate timing information for words, phonemes, and sentences was generated to allow epoching of EEG data to multiple levels of linguistic representation. Log files generated by the task application were used to identify the timing of individual sentences in the task, which were then made temporally precise using a modified version of the Penn Phonetics Forced Aligner (Yuan & Liberman, 2008), which automatically generated Praat TextGrids (Boersma & Weenink, 2013). Automatically generated TextGrids were checked for accuracy at the sentence, word, and phoneme level by authors A. M., N. C., J. H., C. V., V. M., C. T., C. H., and P. P. The first author (G. L. K.) supervised the transcription process and checked the final TextGrids for accuracy before generating event files used in the analyses. Event files containing start and stop times for each phoneme, word, and sentence, as well as information about trial type (perception vs. production; consistent vs. inconsistent playback), were created using the log files and TextGrids. A second set of event files corresponding to the intertrial click sound were generated via a match filter process where the audio signal of the click was convolved with the EEG audio signal to find exact timing matches (Turin, 1960).

To examine the differences between perception and production at the sentence level, sentence-level event files were used to epoch the neural response between −1.5 sec and +3 sec relative to sentence onset, which we quantified as the acoustic onset of the first phoneme of the sentence (Ozker, Doyle, Devinsky, & Flinker, 2022). Epochs ±10 *SD*s from the within-subject mean were rejected. Linear mixed-effects (LME) models were created and assessed using the lmertest package (Kuznetsova, Brockhoff, & Christensen, 2017) in R to determine statistical differences between different task conditions within relevant time windows, specifically the N100 (80–150 msec) and P200 (150–250 msec). The peak amplitudes and latencies of these windows, as well as the peak-to-peak amplitude of the N100 and P200 components, were used as response variables. Latency was calculated as the time at which the largest peak within a time window of interest occurred. LME models were specified using Equation 1:

$$y = X\beta + Zu + \varepsilon \qquad (1)$$

where β represents fixed-effects parameters, u represents random effects, and ε represents residual error. X and Z are matrices of shape ($n \times p$), where n is the number of observations of each parameter and p is the value of the parameter at observation n. In all models, the fixed effect was the response variable of interest (i.e., N100 & P200 amplitude & latency; peak-to-peak amplitude) and subject was used as a random effect. F tests were calculated using Kenward–Roger approximation with n degrees of freedom specified (Kenward & Roger, 1997).

Linear Encoding Model Analysis

Linear encoding models (also referred to as spectrotemporal or multivariate temporal receptive field models in previous literature) were fit to describe the selectivity of the EEG responses to phonological features corresponding to place and manner of articulation (Desai et al., 2021; Hamilton, Edwards, & Chang, 2018; Crosse, Di Liberto, Bednar, & Lalor, 2016; Di Liberto et al., 2015; Mesgarani et al., 2014). This model takes the form of the equation below:

$$\hat{y}_n(t) = \sum_f \sum_{\tau=-0.3}^{\tau=0.5} w(f,\tau)S(f,t-\tau) + \varepsilon \tag{2}$$

where $\hat{y}_n(t)$ represents the estimated EEG signal for electrode n at time t. The stimulus matrix S consists of behavioral information regarding features (f) for each time point $t - \tau$, where τ is the time delay between the stimulus and neural activity in seconds. Features included combinations of binary features for perception, production, consistent playback, and inconsistent playback trials, as well as continuous, normalized EMG activity recorded from auxiliary electrodes, and binary features for the presence of phonological features at each time point (as in Desai et al., 2021; Hamilton et al., 2018; Mesgarani et al., 2014). The "full" model stimulus matrix contained 14 phonological features as well as four binary features encoding trial information (perception, production, consistent playback, inconsistent playback) and normalized EMG activity from facial electrodes for 19 features. These phonological features for place and manner of articulation were identical to those used in previous work (Desai et al., 2021; Hamilton, Oganian, Hall, & Chang, 2021; Mesgarani et al., 2014) and included sonorant, obstruent, voiced, nasal, syllabic, fricative, plosive, back, low, front, high, labial, coronal, and dorsal. Phonemes were coded in a binary matrix where a 1 indicated the onset of a phoneme's articulation via timing information obtained from the TextGrids.

We fit separate models to predict the EEG response in each channel using time delays of −0.3 sec to +0.5 sec, relative to the acoustic onset of the phoneme. This delay range encompassed the temporal integration times to similar responses found in previous research (Hamilton et al., 2018) but with an added negative delay to encompass potential pre-articulatory neural activity (Chartier et al., 2018). Data were split 80–20 into training and validation sets. To avoid overfitting, the data were segmented along sentence boundaries, such that the training and validation sets would not contain information from the same sentence. These segments were then randomly combined into the 80/20 training/validation sets. Weights for each feature and time delay $w(f,\tau)$ were fit using ridge regression on the training set and a regularization parameter chosen by 10 bootstrap iterations, fitting on subsets of the training set. The ridge parameter was selected at the value that provided the highest average correlation performance across all bootstraps. Ridge parameters between 10^{-5} and 10^{5} were tested in 20 logarithmically scaled intervals. Model performance was assessed using correlations between the EEG response predicted by the model and the true EEG response. Significance of these correlations was obtained through a bootstrap procedure with 100 iterations in which the training data were shuffled in chunks to remove the relationship between the stimulus and response but preserve temporal correlations within the EEG signal. Visual inspection of the data revealed two participants (OP4 and OP17) for whom responses showed no discernible receptive field structure even after greatly expanding the range of ridge parameters, motivating their exclusion from the analysis. To investigate the relationship between encoding of phonological features during perception and production, a "task-specific" model was fit (Figure 3A, Figure 5), which contained three sets of phonological features: those that occurred exclusively during production trials, those that occurred exclusively during perception trials, and a combined perception-plus-production set of phonological features identical to those included in the "full" model described above. After fitting this model, we calculated a feature-by-feature correlation for the production-specific and perception-specific feature weights (e.g., correlation of fricative-production with fricative-perception) to investigate how representations of phonological features change between modes of speech (Figure 5B). We also used the production-specific and perception-specific model weights to fit separate predictions of the held-out validation set EEG activity, which were then averaged relative to sentence onset to facilitate comparison to the ERP analysis (Figure 5C).

RESULTS

Topographic inspection of sentence-level ERP activity revealed a frontocentral ROI of nine channels that elicited the strongest response to sentence onset during speech perception and production (F1, Fz, F2, FC1, FCz, FC2, C1, Cz, and C2). This ROI is used in the ERP results, but linear encoding models were fit on all channels for all participants.

ERP Results

After verifying the integrity of the data set, we wished to understand whether and how responses to continuous speech differ for production versus perception and for the consistent and inconsistent playback conditions. Sentence-level ERPs for both perception and production were epoched to the acoustic onset of sentence articulation (the first phoneme in the trial sentence). These ERPs demonstrated a relative suppression of EEG activity in production trials compared with perception trials (Figure 2). The N1 and P2 components are present at the sentence level in both perception and production conditions but

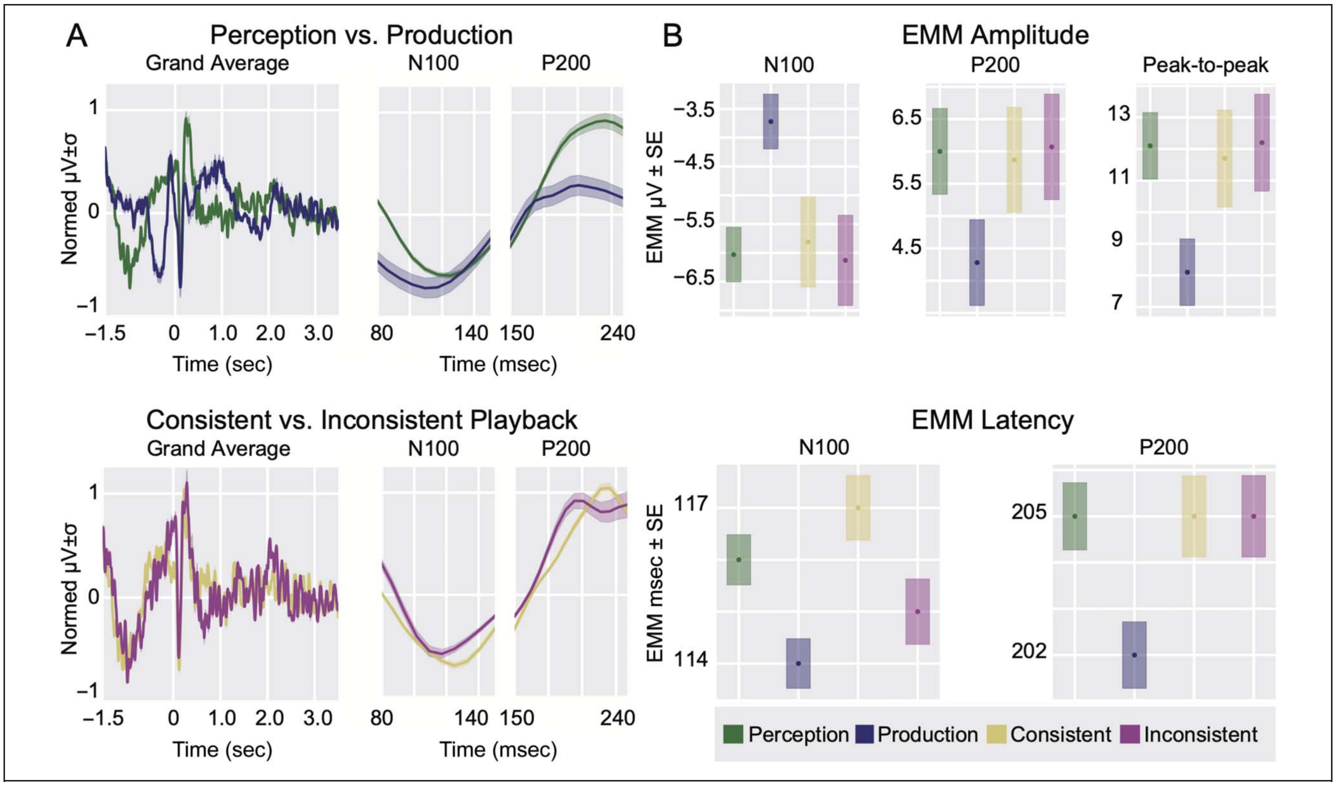

Figure 2. ERPs to sentence onset demonstrate suppression of N1-P2 during speech production. Speech production (indigo) is suppressed relative to perception (green), but no such difference is observable for consistent (yellow) versus inconsistent (magenta) speech perception. (A) Grand average ERPs and N100/P200-windowed ERPs comparing speech production and speech perception (top) and consistent and inconsistent speech perception (bottom). (B) LME model EMMs for the four experimental conditions' amplitudes (top) and latencies (bottom). Shaded area represents standard error.

reduced in amplitude for the production trials. We fit LME models (Equation 1) comparing perception and production in windows of interest (windowed amplitude ~ Condition + (1|Subject) and windowed latency ~ Condition + (1|Subject)). We report the estimated marginal mean (EMM) and standard error of the contrasts between perception and production responses here. We found significantly lower amplitudes for N100 ($\text{EMM}_{\text{perception-production}} = -2.31 \pm 0.15$ µV; $p < .001$) and significantly higher amplitudes for P200 ($\text{EMM}_{\text{perception-production}} = 1.72 \pm 0.15$ µV; $p < .001$) during perception compared with production. This was also in line with increased peak-to-peak amplitude ($\text{EMM}_{\text{perception-production}} = 3.96 \pm 0.15$ µV; $p < .001$) in perception compared with production. In addition, N100 latency was decreased in production compared with perception ($\text{EMM}_{\text{perception-production}} = 1.64 \pm 0.47$ msec; $p < .001$), and similar results were seen for P200 latency ($\text{EMM}_{\text{perception-production}} = 2.75 \pm 0.66$ msec; $p < .001$). Suppression during speech production relative to perception in this task highlights differences in processing internally and externally generated speech.

Next, differences in consistent and inconsistent playback trials were assessed to evaluate the presence or absence of a suppression similar to the one observed between perception and production. Although differences were significant between perception and production trials, the differences between consistent and inconsistent speech perception were less pronounced: LME modeling (Window ~ Condition + (1|Subject)) did not reveal a significant difference in N100 ($\text{EMM}_{\text{consistent-inconsistent}} = 0.31 \pm 0.20$ µV; $p = .12$) and P200 ($\text{EMM}_{\text{consistent-inconsistent}} = -0.20 \pm .22$ µV; $p = .37$) amplitudes across this contrast. However, peak-to-peak amplitude ($\text{EMM}_{\text{consistent-inconsistent}} = -0.52 \pm 0.24$ µV; $p = .03$) and N100 latency ($\text{EMM}_{\text{consistent-inconsistent}} = 1.50 \pm 0.66$ msec; $p = .02$) differed significantly between consistent and inconsistent trials, with an earlier response to inconsistent compared with consistent playback. P200 latency did not differ significantly ($\text{EMM}_{\text{consistent-inconsistent}} = 0.18 \pm 0.91$ msec; $p = .84$). Because consistent and inconsistent perception trials were split into blocks during the task, an oddball response was not elicited for the unpredictable stimuli. To further investigate the significance of peak-to-peak amplitude and N100 latency between consistent and inconsistent perceptual stimuli, a series of Wilcoxon signed-ranks tests with Benjamini-Yekutieli correction (Benjamini & Yekutieli, 2001) comparing N100-P200 peak-to-peak amplitude and N100 latency on a within-subject basis were performed. These significance tests revealed only three individual participants that demonstrated a significant suppression between consistent and inconsistent speech perception (OP1 $p = .02$; OP7 $p < .001$; OP21 $p = .0002$), and only two participants with a significant difference in N100 latencies (OP1 $p = .004$; OP19 $p = .04$). This within-subject analysis suggests the significance of peak-to-peak amplitude and N100 latency observed in the LME results is caused by outlier participants rather than a generalizable effect. Overall, differences within consistent and inconsistent perception trials were less pronounced than the differences between perception and production trials. These minor differences between expected and unexpected speech perception suggest that SIS is not fundamentally linked to general forward modeling of speech production. In other words, feedforward processing of speech perception and feedforward processing of speech production reflect different neural mechanisms.

Linear Encoding Model Results

While our ERP results provide insight into the timing and magnitude of differences in responses during perception and production, they do not provide information regarding any potential differences in responses to specific speech features or content. Furthermore, ERP analyses are constrained by the need to average many trials that are time-locked to a particular event (Luck, 2014). Thus, ERP analyses may not be as sensitive to uncovering differences outside of the onset of the sentence, or for specific phonological features within continuous speech. To address this limitation, we performed additional analyses where we fit linear encoding models for continuous production and perception (Equation 2). These analyses are powerful in that they allow for investigation of continuous, natural speech without the need for trial averaging. Although we could perform a phoneme-by-phoneme ERP analysis to show task-related differences across the sentence, such an analysis would suffer from an inability to account for coarticulation or other temporal correlations of activity. The linear encoding model regression weights are calculated for multiple time delays simultaneously, allowing the model to account for activity in response to combinations of features across time (Theunissen, Sen, & Doupe, 2000). They also allow us to further probe specific differences (or lack thereof) in tuning across our different task conditions.

Model performance was evaluated by calculating the linear correlation coefficients (r) between the EEG response predicted by the model and the actual response for held out data not used to train the model. We also probed the importance of individual features on model performance by ablating specific features from the stimulus matrix S and observing the change in correlation coefficients between ablated and full models. Similar variance partitioning methods have been used to uncover the unique variance explained by particular features (Hamilton et al., 2021; De Heer, Huth, Griffiths, Gallant, & Theunissen, 2017). For example, if a model that omitted normalized EMG predicted the neural response less accurately, the interpretation is that EMG contains important information for accurately modeling EEG activity. For each task-related feature in the “full” model (14 phonological +4 task features; Figure 3C), we fit a separate model omitting that

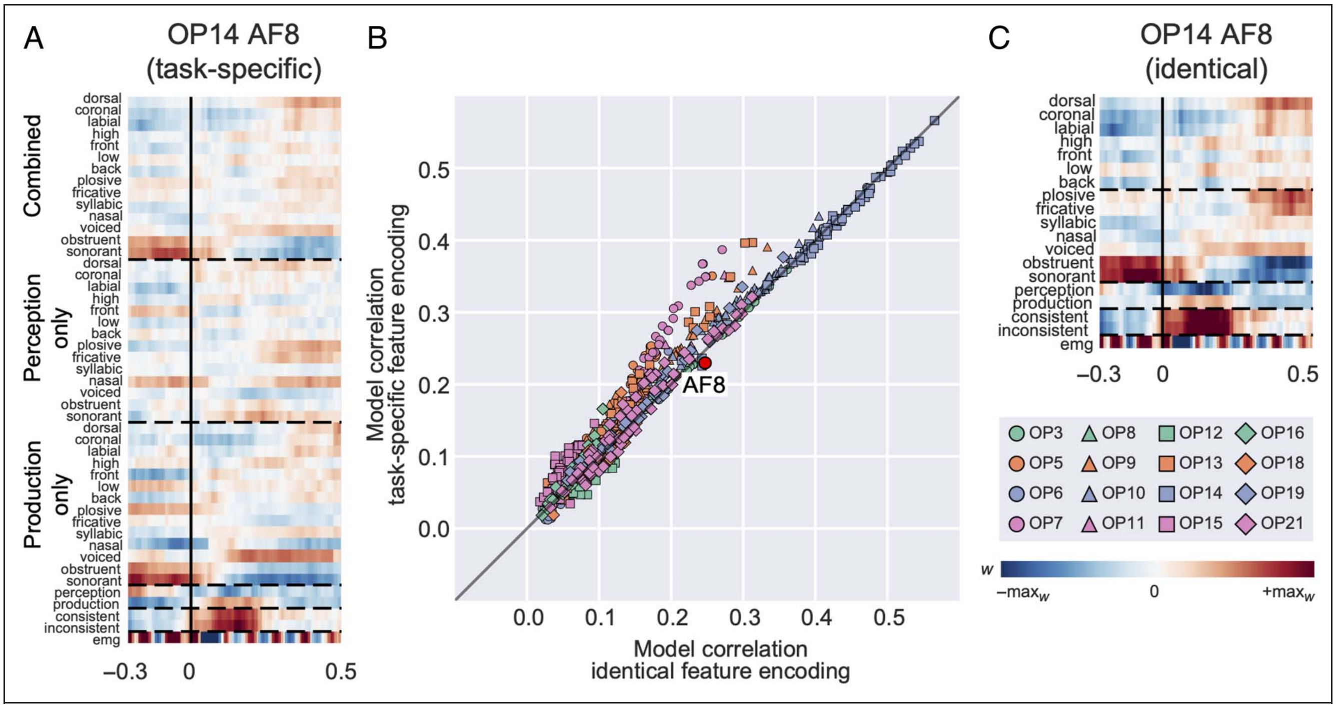

Figure 3. Separating phonological feature encoding by modality of speech improves model performance. (A) Temporal receptive field for an individual electrode with stimulus characteristics divided by task condition (i.e., perception vs. production). (B) Scatter plot of channel-by-channel correlation coefficients between two compared models. Color and markers are used to denote individual participants. Diagonal black line represents unity (equal model performance). (C) Temporal receptive field for an individual electrode with stimulus characteristics identical across task condition.

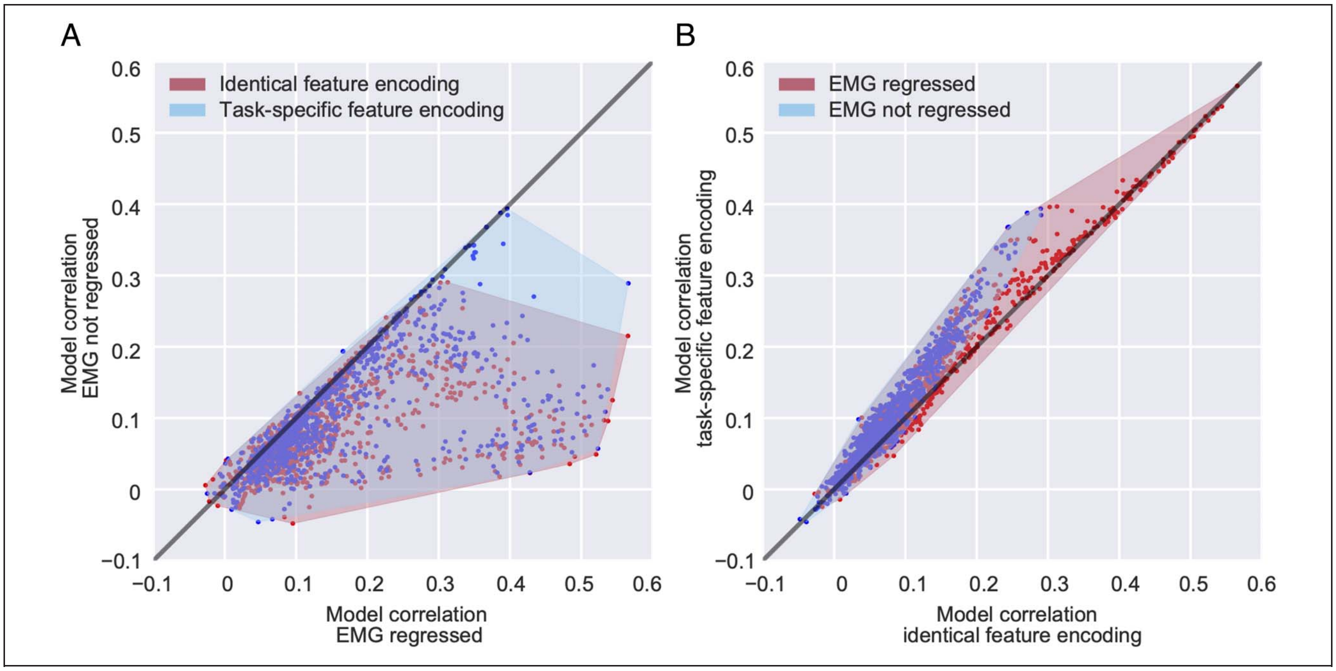

Figure 4. Including EMG as an encoded feature in linear models greatly improves their performance, as well as the stability of phonological feature encoding between perception and production. (A) Individual electrodes' correlation coefficients with held-out neural response within models that do contain an EMG regressor (*x* axis) and those that do not (*y* axis), for models that separate phonological feature tuning by task modality (blue) and models that do not (red). Diagonal black line represents unity. Shaded area is the convex hull of points within each group to show overall trends. (B) Individual electrodes' correlation coefficients with held-out neural response within models that differentially encode phonological features according to modality of speech (*y* axis) and those that do not (*x* axis), in the presence (red) or absence (blue) of information about normalized EMG activity recorded from auxiliary facial electrodes. Diagonal black line represents unity. Shaded area same as (A). When EMG was regressed, more points lie along the unity line, indicating similar phonological feature tuning and that EMG may be captured in the different phonological features when it is not available as a regressor.

feature. Lastly, one model had two additional sets of phonological features (i.e., 14 phonological features during production + 14 phonological features during perception + 14 phonological features in either condition + 4 task features; Figure 3A). These were split by modality to observe if phonological feature tuning changed between perception and production. We call this model the "task-specific" encoding model, which is in comparison to the "identical" encoding model in which phonological feature tuning is assumed to be the same across all conditions, with only a baseline change fit by the two condition features (perception and production). The EMMs of the contrast in correlation coefficients between models were evaluated via LME modeling with subject and channel location as random effects ($r \sim$ Model + (1|Subject) + (1|Channel)). Separating phonological tuning by the modality of speech (i.e., perception vs. production) had a significant effect on model performance (EMM$_{\text{identical-separate}}$ $r = -0.014 \pm 0.003$; $p < .001$), such that separating phonological feature tuning during production from phonological feature tuning during perception improved the model's ability to predict the held-out neural response (Figure 3B). This result, which was contrary to our initial hypothesis, suggested that phonological feature encoding differs during speech perception and production. However, because of the influence of EMG artifact during speech production, speech perception in this task is a combination of sensory and motor responses, whereas speech perception in this task is purely sensory, which may explain the difference in the models presented in Figure 3.

Although we utilized methods to correct for EMG artifact that have been previously demonstrated in the literature to be successful (Riès et al., 2021; Chen et al., 2019; de Vos et al., 2010), there is no definitive way to rule out residual EMG given the lack of ground truth in the sources that contribute to the electroencephalogram. As a result, we further explored the influence of EMG artifact on model performance by fitting linear encoding models that included normalized EMG activity recorded from auxiliary facial electrodes in tandem with the EEG as a regressor. Models that include or exclude the auxiliary EMG but are otherwise identical in their stimulus matrices were compared in an ablation-based approach to explore the contribution of specific features to model performance (Ivanova, Hewitt, & Zaslavsky, 2021). Linear correlation coefficients were compared using an LME model identical to the model used for comparing the "identical" versus "task-specific" models described above. The inclusion or exclusion of normalized EMG in the stimulus matrix significantly affected model performance regardless of whether phonological features were task specific ($p < .001$) or identical ($p < .001$). Including information about normalized EMG activity recorded from auxiliary facial electrodes improved model performance (Figure 4A) as shown by the greater number of channels below the unity line. On an individual participant basis, all but two participants (OP6 and OP16) showed a significant difference in model performance across the inclusion or omission of normalized EMG activity as a stimulus feature as assessed by Wilcoxon signed-ranks test. When comparing the relative difference between "identical" and "task-specific" models (Figure 3) in the presence or absence of an EMG regressor, models including an EMG regressor showed less of a difference in performance between methods of phonological feature encoding, suggesting that residual EMG decreases the stability of phonological feature tuning across modalities of speech (Figure 4B). A verification of artifact removal in the context of the ERP results reported above is provided in the Appendix.

Although the linear encoding model results occur on a phoneme-by-phoneme timescale (cf. the sentence-level ERPs presented in Figure 2A), the EEG data used to fit the models were collected during a task that elicited SIS. Thus, we sought to identify any reduction in phonological feature response between production-specific and perception-specific feature weights in an effort to link the linear encoding model results to the ERP analysis presented earlier in this article. Feature-by-feature, we

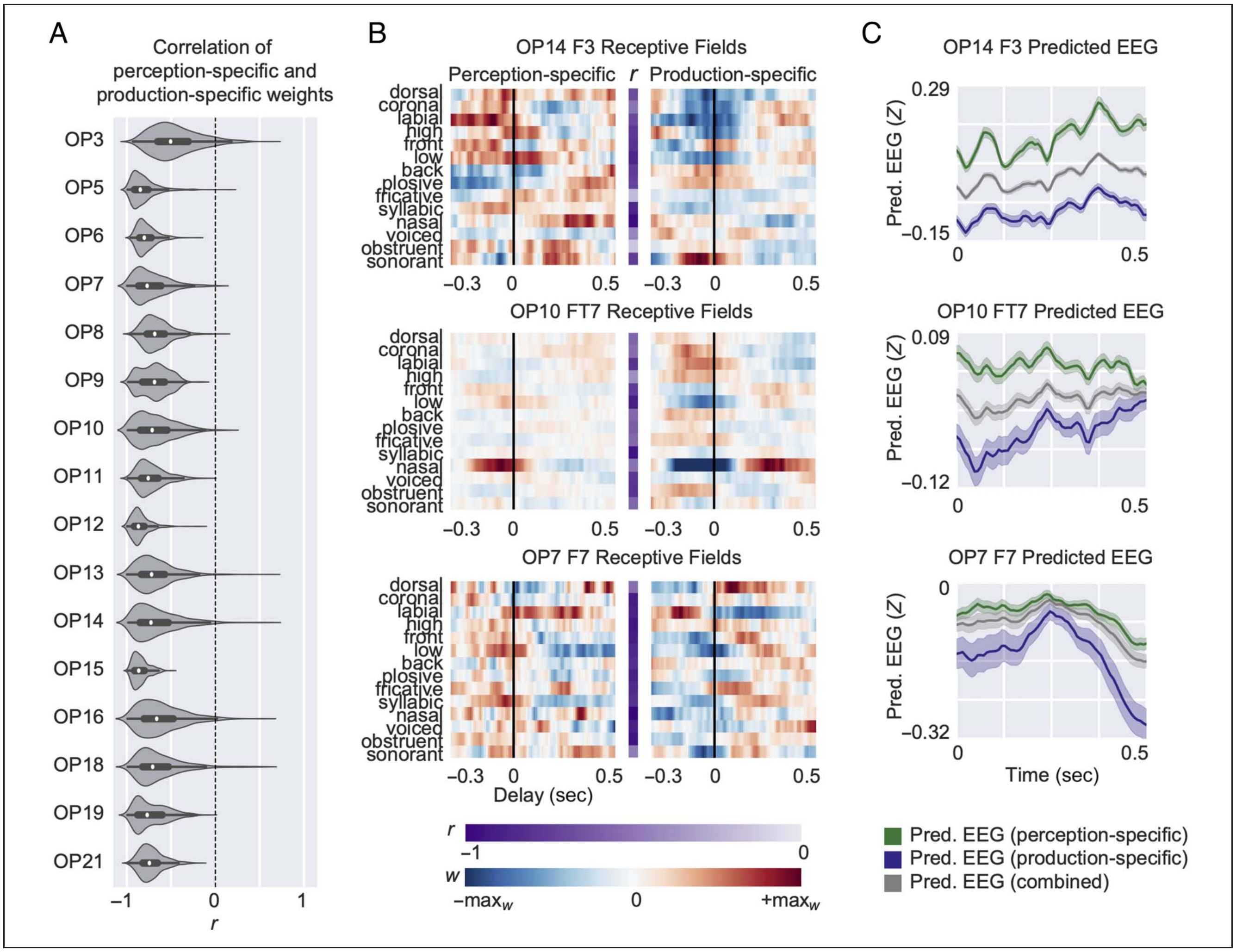

Figure 5. Production-specific and perception-specific phonological feature weights are strongly negatively correlated with each other, suggesting a suppressive relationship. (A) Violin plot showing the distribution of channel-by-channel, feature-by-feature correlation coefficients between phonological features specific to perception and phonological features specific to production, separated by individual participant. Thick line in violin interior represents range of Quartiles 1–3. Density of plot (violin width) scaled by individual participant. (B) Temporal receptive fields for three individual electrodes. Phonological feature weights taken from task-specific model separated into perception-specific (left) and production-specific (right) receptive fields. Center grayscale column represents the correlation of each row of weights between the two receptive fields. (C) Predicted EEG activity for the held-out validation set as predicted by the perception-specific (green), production-specific (indigo), or combined (gray) phonological feature weights. Electrodes OP14 F3 (predicted vs. actual EEG $r = .53$; $p < .001$) and OP10 FT7 (predicted vs. actual EEG $r = .42$; $p < .001$) exhibit similar model performance between task-specific and identical phonological feature encoding models (i.e., lie along unity line of Figure 3B), whereas electrode OP7 F7 (predicted vs. actual EEG $r_{\text{task-specific}} = 0.37$; $r_{\text{identical}} = 0.24$; $p < .001$) exhibits diverging model performance between models. Overall, predicted EEG based on the production-specific weights was lower in amplitude than predicted EEG based on the combined or perception-specific weights. All linear encoding models presented in this figure are from the task-specific model that included an EMG regressor.

calculated the correlation coefficient (*r*) of the production-specific and perception-specific weights of the task-specific encoding model (Figure 3A) and observed a strong negative correlation between the production and perception-specific weights (Figure 5A).

Trial-specific stimulus features were also ablated to assess their contribution to model performance. Omitting trial modality (i.e., whether a phoneme was produced or perceived) did not significantly affect the linear regression model's ability to predict the held-out neural response (p = .65). Similarly, ablating information about whether the perception trials were consistent or inconsistent with their preceding production trials did not affect model performance (p = .56). If the EMG regressor is removed in conjunction with trial-specific features, the differences in model performance when trial modality is included or ablated are less profound but still nonsignificant (p = .23). When ablating playback consistency, no changes are observed in significance between inclusion (p = .56) and omission (p = .54) of an EMG regressor, which is expected considering this contrast is constrained to perception trials where EMG associated with articulation is absent from the response. The ablation of consistency contrast not affecting model performance is in line with the ERP results presented above (Figure 2). However, ablating trial modality (i.e., perception vs. production) not affecting model performance is incongruent with the ERP results, for which a stark contrast between perception and production were observed. The difference in time frame between the ERP analysis (sentence level) and the linear encoding model analysis (phoneme level) may explain the difference between the ERP and linear encoding model results. In other words, sentence-level processing of speech perception and production may involve different neural mechanisms, but at an individual phoneme level, the mechanisms are shared between perception and production. Alternatively, the incorporation of the EMG regressor may be delineating perception and production in the model, making explicit information about trial modality, effectively marking the explicit inclusion of trial type in the stimulus matrix redundant. This explanation is supported by the observation that omission of an EMG regressor substantially impacted model performance.

Taken together, the linear encoding model results suggest that linguistic abstractions remain invariant during speaking and listening. Similar encoding of phonological features between these modes of speech after EMG regression suggests the amplitude reduction corresponding to SIS is not explicable by a difference in linguistic abstraction, constraining it to endogenic sensorimotor processes. However, a feature-by-feature correlation shows an inverse relationship between the encoding of phonological features during speaking and listening, demonstrating a suppressive relationship between speaking and listening can be observed on an individual phoneme timescale throughout sentences that exhibit SIS at sentence onset. Methodologically, the linear encoding model results show that regressing EMG activity recorded from auxiliary electrodes during the task is an informative characteristic of the stimulus in the context of modeling neural responses to speech. Including information about trial type (perception vs. production, consistent vs. inconsistent playback) was less informative when EMG was included as a regressor, potentially the result of fundamental differences in expected residual EMG between articulating speech and passively listening. EMG regression also reduces phonological feature tuning changes across modality of speech, suggesting residual EMG artifact in the postprocessed signal is responsible for changes in phonological feature tuning, as well as motivating auxiliary EMG recordings as a safeguard against residual EMG in the postprocessed response.

DISCUSSION

Summary

The results presented in this study demonstrate a difference in EEG responses to perceiving and producing naturalistic stimuli. At the sentence level, a suppression of early auditory components N100 and P200 was observed in speech production relative to perception. These findings are in line with previous literature on SIS and auditory processing more generally. The N100 (and its MEG equivalent N100m) have been theorized as a neural indicator of the efference copy, and its suppression has been demonstrated for internally generated speech compared with externally generated speech (Behroozmand & Larson, 2011; Martikainen et al., 2005). The P200 is less directly associated with SIS, with limited studies linking it directly to feedback perturbation (Brumberg & Pitt, 2019; Behroozmand & Larson, 2011), but it is commonly paired with the N100 in speech perception studies to form the N1-P2 complex (Lightfoot, 2016). In contrast with the suppression observed between perception and production trials, differences between EEG responses to consistent and inconsistent perceptual trials were minor. Even if major differences between these trials were found, our experimental manipulation of playback consistency would be unable to link forward modeling during speech perception to feedforward control during speech production. The limitations of this manipulation in the context of this research question are discussed below.

To investigate whether linguistic abstraction into phonological features persists during SIS, in neural responses to these two modes of speech, we fit linear encoding models describing neural activity as a function of different stimulus features. These features allowed us to test different hypotheses about changes in phonological tuning at the individual feature level versus overall baseline changes during perception and production. Performance of these models were evaluated by how well the weights of the models correlated with held-out EEG response. Differentially encoding phonological features during perception and production in the stimulus matrix yielded higher model performance; however, residual EMG artifact may be driving performance improvements in the differential phonological features model, considering the inclusion of normalized EMG recorded from facial electrodes substantially improved model performance. As EMG activity is expected to disproportionately affect speech production because of articulatory movement, residual EMG that is unaccounted for with an additional regressor may be providing the model with a clear contrast between the perception and production conditions that is spuriously encoded in the separation of phonological features across modes of speech. Accordingly, the inclusion of the EMG regressor reduces the variance in phonological feature encoding between perception and production by accounting for uncorrected EMG artifact in a separate regressor. Thus, we conclude that phonological feature encoding is a shared representation during speaking and listening. Despite similar ability (after regressing EMG) to predict held-out EEG as models that do not separate phonological features into whether they occurred during perception or production, models that do separate phonological features show a strong negative correlation between task-specific phonological features (Figure 5). Taken together, the ERP and linear encoding model analyses extend our understanding of SIS by scaling the phenomenon into a more naturalistic context and exploring the interaction of SIS and higher-order linguistic abstractions that take place during speech perception and production.

Potential Mechanisms of SIS

Previous literature comparing speech production and suppression has identified a neurophysiological effect dubbed speaker-induced suppression, where internally produced stimuli generate less of a change in neural activity than externally produced stimuli. This study sought to replicate this effect in a more naturalistic setting, as many studies of SIS use low-level acoustic stimuli such as pure tones (Martikainen et al., 2005) and single vowels (Niziolek et al., 2013; Heinks-Maldonado, Nagarajan, & Houde, 2006; Houde et al., 2002), whereas many neurolinguistic studies now use more naturalistic stimuli such as podcasts (Goldstein et al., 2022; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016), audiobooks (Herff et al., 2015), and movie trailers (Desai et al., 2021) in an effort to better capture how speech and language are used in daily life (Hamilton & Huth, 2020). We were able to demonstrate SIS at the sentence level, which is comparatively much more naturalistic than the lower-level characteristics of speech used in previous studies of SIS.

SIS emerges from neural synchrony between expectations about utterance content generated before articulation and the sensorimotor/auditory feedback generated during articulation. More specifically, the degree of suppression during speech production is linked to how well the production of a speech token matches with a canonical sensory goal for that token. For example, for vowel production, a prior study associated the degree of suppression on a per-utterance basis with the proximity of the first and second formants to the averaged acoustic center of that vowel (Niziolek et al., 2013). That study demonstrated that the efference copy, a feedforward expectation about the content of the upcoming auditory stimulus only generated during internally produced speech, contains goal-oriented information. A hypothesis put forth by the authors is that the efference copy, although itself a precise encoding of motor commands, loses its precision in sensory cortex in favor of invariant encoding of information.

The extent of this goal-based representation (i.e., Does it extend to higher levels of linguistic abstraction?) motivated the analysis of phonological feature tuning during SIS. During speech production, linguistic representations are used during pre-articulatory planning, meaning the invariant representations are already available to the language network, negating the need for additional processing costs associated with segmentation of continuous acoustic information, a necessary component of properly perceiving speech. Previous research has shown that electrodes in sensory cortex are preferential to specific classes of phonological features (Mesgarani et al., 2014), which motivated their investigation in the present study. Although the degree of SIS remains sensitive to subphonemic changes in auditory feedback, are the invariant representations used in sensory cortex suspect to differences between speech perception and production? We expanded on the N100 suppression observed in the ERP by ablating specific stimulus characteristics from linear encoding models and observed how the absence of a specific aspect of the stimulus affected the model's ability to predict the neural response. Using the same data in which SIS has been identified per the ERP results, a concurrent reduction in phonological feature response was identified via the linear encoding model approach (Figure 5). This suggests that neural activity at multiple levels of representation (sensorimotor activity per SIS, phonological/linguistic representation per linear encoding model) was suppressed during speaking when compared with listening. However, the observation that regressing EMG activity increased stability of phonological feature tuning across modalities, coupled with the ablation of information about stimulus modality, suggests that residual EMG artifact may be driving the differences between overall model performance (Figure 4). Nevertheless, the structure of the receptive fields themselves was relatively consistent (albeit inverted)—that is, the brain does not shift toward representing different phonological features during perception and production (Figure 3A, 5B). Although the present study showed consistent feature representations between speaking and listening despite magnitude changes within those representations, one previous study did identify changes in feature representation between speaking and listening in the motor cortex (Cheung et al., 2016). In that study, motor electrodes clustered according to place of articulation during speech production, while during passive listening, they clustered according to manner of articulation. However, the authors mention that it is unclear whether these differences in feature representation are the result of a single intracranial electrode recording from two different populations of neurons (one sensory and one motor) or whether the same population changes its representation depending on task. Because our study used scalp EEG electrodes, we were unable to distinguish between sensory and motor areas with our recorded level of spatial precision.

Pre-articulatory Readiness Potential in Speech Production

In our ERP analysis, we observed a positive deflection in the grand average ERP (Figure 2A) that began ~200 msec before articulation and peaked ~100 msec before articulation present in the speech production trials. We believe this activity to be related to feedforward linguistic and motoric preparation that must take place before articulation. Before articulation, a communicative desire must be morphologically, syntactically, and lexically encoded before it is transformed into a motor program for the speech articulators (Flinker et al., 2015; Tourville & Guenther, 2011; Levelt, 1993). The exact pre-articulatory stages of speech production are difficult to dissociate with this task, as there is no epoched timing information available as to when these processes occur in a naturalistic context; however, the presence of this pre-articulatory activity exclusively during speech production motivates these stages as an explanation. Prestimulus activity was also observed in the grand average during perception trials in the form of positive activity starting at −600 msec and peaking at stimulus onset. This activity may be related to predictive components of speech perception, as feedforward processing is an important aspect of successful speech perception (Hamilton et al., 2021; Heald & Nusbaum, 2014; Poeppel & Monahan, 2011). This speculation is supported by the structure of the task allowing participants to anticipate when they would hear a sentence; however, this task was not operationalized in a way that allows a more granular analysis of this phenomenon. Notably, for both perception and production, the polarity of the prestimulus activity was inconsistent from subject to subject. This internal inconsistency suggests the activity is not related to previously described ERP components (e.g., readiness potential/Bereitschaftspotential) as these components have a canonical negative polarity (Jahanshahi & Hallett, 2003; Yoshida et al., 1999; Wohlert, 1993). An alternative explanation for pre-articulatory activity in this task is this activity is reflective of residual uncorrected EMG; however, the integrity of task-related neural components suggests any EMG activity capable of producing such a large deflection would not be present in the corrected data (see Appendix).

Influence of EMG Artifact

In any noninvasive neuroimaging study of speech production, movement artifacts caused by articulation are a concern to the integrity of the data. Traditionally, EEG analyses of speech production have sidestepped addressing EMG by requesting participants "imagine" speech while not moving the articulators or shifting the analysis window to a time outside when articulatory movement is occurring. However, there have been recent successful attempts at directly analyzing the window of overt articulation up to the phrase level (Riès et al., 2021). In that experiment, stimuli consisted of four-word tongue twisters. We extend these results by scaling up to the sentence level with evoked responses to speech appearing relatively cleaned of EMG artifact as evidenced by the integrity of the N100 and P200 components. Because of the success of these prior studies in analyzing event-related EEG data during overt speech production, we did not provide further corroboration of the artifact correction techniques used as a primary result of our study. However, because this study uses a more continuous speech stimulus than the prior studies described above, we provide an investigation into the efficacy of our artifact correction techniques in the Appendix.

One reason to assume that residual EMG is affecting the results is the differing performance of encoding models that do or do not regress EMG (Figure 4). Models that accounted for EMG as a stimulus characteristic on the whole outperformed models that did not, which means there is variance remaining in the postprocessed data that is well explained by EMG activity. The inclusion of an EMG regressor was only made possible by recording facial muscle activity using auxiliary electrodes in conjunction with the EEG, akin to how EEG researchers will record auxiliary VEOG and HEOG to assist with artifact correction. Although previous research has demonstrated blind source separation-based artifact correction techniques are sufficient in correcting EMG artifact for ERP analysis, the substantial difference in model performance when this normalized EMG activity was ablated from the stimulus matrix leads us to strongly recommend the use of auxiliary EMG recordings to any researchers who wish to fit similar linear encoding models to speech production data. Furthermore, we only recorded single-channel EMG, whereas there are a plethora of facial muscles that contribute to EMG artifact in the electroencephalogram. It is possible that including activity from multiple auxiliary channels as a regressor in linear encoding models would further improve their performance, but future research is needed to substantiate this claim.

There are several reasons we do not believe the residual EMG in our response nullifies the interpretation of this study's results. First, the integrity of purely auditory responses is preserved after post-processing as evidenced by evoked responses to intertrial click tones, which suggests the evoked responses seen at the sentence level are not false positives caused by EMG artifact (see Appendix). Second, despite the contribution of EMG to linear encoding models, we observe strong phonological feature tuning consistent with previous research (Desai et al., 2021; Hamilton et al., 2021). Third, including EMG as a regressor in linear encoding models ensures that phonological feature tuning (or a similar feature space of interest) is not obscured or affected by muscle artifact. Lastly, evoked responses to sentence onset contained robust N100 and P200 components that would not be visible in the presence of substantial noise from EMG.

Perceptual Stimulus Consistency

A manipulation of whether the perceptual trials immediately followed the production trials from which they were generated (consistent) or not (inconsistent) was included in the present study to assess the hypothesis that SIS is associated with general feedforward auditory processing and not an intrinsic characteristic of corollary discharge during speech production. Differences between consistent and inconsistent perceptual trials were small, with only three individual participants demonstrating a significant difference. Many studies have demonstrated consistency in the task structure, such as our manipulation of feedback match/mismatch, can affect behavioral results, but there is no corroborated link between these results and SIS (Lester-Smith et al., 2020; Mollaei, Shiller, Baum, & Gracco, 2016; Gonzalez Castro, Hadjiosif, Hemphill, & Smith, 2014; Jones & Munhall, 2000). Studies that show an expectancy effect in EEG identify differences in later components such as the N400 and P600 (Goregliad Fjaellingsdal et al., 2020) as well as earlier components such as mismatch negativity (Bendixen, Scharinger, Strauß, & Obleser, 2014; Hawco et al., 2009; Näätänen, Paavilainen, Rinne, & Alho, 2007) and the N1 (Astheimer & Sanders, 2011), of which the latter has been posited as a neural biomarker of the efference copy (Behroozmand & Larson, 2011; Heinks-Maldonado et al., 2007; Martikainen et al., 2005). The present study's manipulation of playback consistency, although hypothesized to elicit a similar response to expectancy effects present in the EEG literature, failed to do so. In part, this may be because of the task's block design, which was used to explicitly avoid eliciting an oddball response. Our study presented the consistent and inconsistent perceptual trials in blocks of 50 trials each, which means participants could identify when perceptual stimuli would be inconsistent with the preceding trial, a fundamental difference from the oddball tasks where deviant stimuli are presented randomly. We additionally chose not to present inconsistent stimuli in an oddball fashion because our perceptual stimuli were generated from the recorded productions of the participant. Thus, to generate the full range of inconsistent perceptual stimuli in our task, a full block of production trials is needed, and collecting this as a baseline before introducing oddball inconsistent stimuli would greatly extend the time of our recording sessions, and we judged more repetitions of each condition to be more important to our research questions. The block design of the task may also cause listeners to adapt to the randomly shuffled perceptual stimuli over the course of the block.

In addition, other top–down influences on auditory processing and a non-uniformity of suppression across cognitive processes makes the initial research question regarding a connection between forward modeling of auditory stimuli in perception and production difficult to investigate with this experimental paradigm. In other words, the block-based experimental design of our playback consistency manipulation imposes additional limitations on its interpretability. As mentioned above, we opted to use a block-based design to minimize acoustic differences between experimental conditions and to maximize the number of repetitions in each condition without extending the length of the task for participants. A similar experiment in which all inconsistent playback trials were interspersed randomly among consistent playback trials would facilitate a comparison to conventional "oddball" studies of predictability in the EEG literature. Designing a study in this way may also minimize the influence of potential extraneous top–down manipulations on auditory processing. Top–down manipulations of expectations about perceptual content were a variable of interest for the present study, which is why we avoided an oddball design. The naturalistic design of our stimuli introduces many potential top–down processes, not just the forward modeling of perceptual trial content that we sought to investigate. For example, participant engagement with the stimuli can affect the degree of SIS observed. Prior studies have found reduced N100 amplitude during active listening when compared with passive listening (Brumberg & Pitt, 2019; Houde et al., 2002). Our participants were not instructed to actively listen and were not required to make any responses concerning the playback condition, meaning within-subject differences in attentive listening were left up to the independent engagement of the participant with the task. Speech production requires active listening to monitor for potential errors in the speech signal that can subsequently be corrected, whereas speech perception does not. Beyond multitudinous top–down effects on perception during our task, suppression can also emerge from many cortical sources. Thus, linking any suppression observed during this perception-only manipulation to the cross-modal suppression observed between speaking and listening is not trivial. Although we maintain there may be links between general forward modeling during speech production and expectations about sensory consequences during passive listening, the manipulation used in our playback trials is not sufficient to elucidate any of those links. In the retrospective context of our mostly null playback consistency manipulation, more attention to the precise manipulation of this contrast in future studies may yield more informative results, including potentially exploring differences in the N100 between active and passive speech perception trials.

Conclusion

The current study uses naturalistic sentence-level speech perception and production to examine how EEG responses to speech differ across modality of speech and how responses to playback of speech are influenced by expectations about the content of playback. After correcting for EMG artifact in the data via CCA, we demonstrate SIS in production relative to perception as demonstrated by amplitude and latency of the N1-P2 complex epoched to sentence onset. No major findings were observed for consistent versus inconsistent playback contrast in the perception trials. Next, we examined phonological feature tuning across these behavioral conditions via a series of linear encoding models. A difference in tuning was observed such that separating phonological feature encoding across perception and production improved model performance, but these gains were attenuated by the inclusion of normalized EMG as a regressor, suggesting that residual EMG in the model decreases stability of phonological tuning across modalities of speech. With the inclusion of an EMG regressor, stable phonological feature tuning between production and perception confirms the confinement of SIS to lower levels of sensorimotor processing. This research hopes to illuminate differences in electrophysiological responses to perception and production and motivate future study of naturalistic speech production with noninvasive EEG.

APPENDIX: VALIDATION OF EMG ARTIFACT CORRECTION

We used a blind source separation technique based on CCA to identify and correct for EMG artifact (De Clercq et al., 2006). This approach has been successful in removing EMG artifact associated with articulation in previous EEG studies of overt speech production (Riès et al., 2021; Riès et al., 2013; de Vos et al., 2010); our approach is described in the Methods section of this article. These prior studies have focused on speech production at the word or phrase level whereas our study focuses on sentence-level speech production, a more naturalistic form of speech that could potentially elevate the risk for EMG contamination of the data. As a safeguard, we recorded EMG via auxiliary facial electrodes during the task (Figure 1B). Including these recordings as a regressor in linear encoding models reduces the influence of residual EMG artifact on the response (Figure 4); however, for the ERP analyses, external validation of the data set's integrity was necessary to deem it suitable for analysis.

To accomplish this, we compared EEG responses to the task before and after CCA artifact correction. Responses were epoched to the acoustic onset of the first phoneme of the sentence (as in Figure 2A), as well as two non-task-related events: the acoustic onset of the intertrial click tone and peaks in the auxiliary EMG electrode activity as detected by the MNE-python function *mne.preprocessing.create_eog_epochs* (Figure A1A). Significance between pre- and post-CCA-corrected epochs was assessed via LME modeling using a technique similar to the one described in the above article: a fixed effect of trial type and a random effect of subject (RMS Difference ~ Condition + (1|Subject). However, instead of raw voltage values, difference waves between uncorrected and CCA-corrected activity were calculated at each epoch by subtracting the root-mean-square of the two responses averaged across channels, then averaging across the first 300 msec of activity relative to the epoch of interest. Although polarity is important for interpreting EEG components such as N100 and P200, we opted to make no assumptions about the polarity of potential EMG artifact by using the root-mean-square of the response. If EMG were successfully removed from our data, epochs associated with EMG activity (speech production and peak auxiliary activity) should show a larger difference wave between uncorrected and CCA-corrected activity than epochs unassociated with EMG activity (speech perception and intertrial clicks).

Linear-mixed effects modeling (Equation 1) provided a confirmation that EMG was successfully removed from the data set using CCA while preserving the integrity of the neural response (Figure A1B). We report the EMM and standard error of the difference waves here. Epochs associated with articulatory activity showed a large difference before and after CCA artifact correction ($\text{EMM}_{\text{Production}} = 58.40 \pm 78.6\ \mu\text{V}$; $\text{EMM}_{\text{Aux Peak}} = 87.7 \pm 79.4\ \mu\text{V}$), whereas epochs associated with passive listening showed a small difference before and after CCA artifact correction ($\text{EMM}_{\text{Perception}} = 12.1 \pm 79.2\ \mu\text{V}$; $\text{EMM}_{\text{Click}} = 19.8 \pm 77.5\ \mu\text{V}$). The difference in how these epoch types were affected by CCA was further corroborated by the effect sizes of the LME model's contrasts, calculated as Cohen's *d* (Cohen, 2013). Contrast between epochs that both involve articulation or both involve passive listening were small ($\Delta_{\text{Aux Peak-Production}}\ d = 0.038, p = .39$; $\Delta_{\text{Click-Perception}}\ d = 0.01, p = .96$), whereas contrasts between epochs that differed in their expected contamination with EMG activity were large ($\Delta_{\text{Click-Aux Peak}}\ d = -0.089, p < .001$; $\Delta_{\text{Click-Production}}\ d = -0.05, p = .07$; $\Delta_{\text{Aux Peak-Perception}}\ d = 0.1, p < .001$; $\Delta_{\text{Perception-Production}}\ d = -0.06, p = .06$).

Our validation techniques suggest that CCA is a sensitive and specific method for correcting EMG activity in our data set. Although these results are promising, an important caveat is that there is no guaranteed method of confirming an artifact technique is both successful (no Type I error) and accurate (no Type II error); EEG has no "ground truth" for source localization (Bradley, Yao, Dewald, & Richter, 2016). This caveat motivates the use of external validation techniques described here, but also imposes a fundamental limitation on all EEG studies which employ artifact correction techniques. We argue for the integrity of our results despite this limitation, and we encourage those interested in using artifact correction techniques to study naturalistic speech production via EEG to do so.

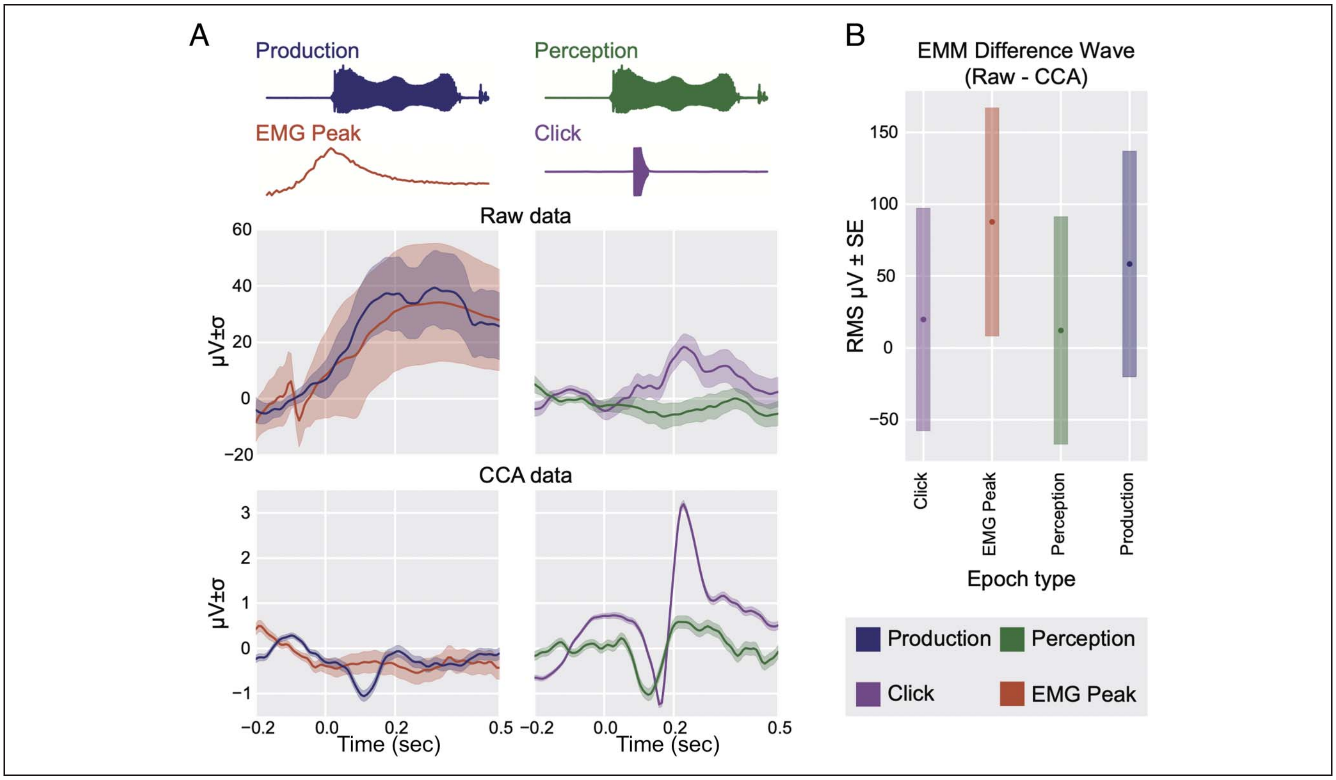

Figure A1. Comparison of EEG activity before and after EMG artifact correction. (A) Stimuli (top) and grand average ERP of raw data (middle) and CCA-corrected data (bottom) relative to displayed stimuli. Grand average plots are separated by the epochs' anticipated level of contamination with EMG artifact. Left panels show epochs that are anticipated to be contaminated because of their association with articulation. Right shows epochs that are anticipated to contain relatively less EMG artifact because of their association with passive listening; however, jaw clenching during passive listening means these data cannot be assumed to be EMG free. (B) LME model EMMs for the RMS amplitudes of 0–300 msec raw-CCA difference waves for each of the four epochs of interest. Shaded area represents standard error. A value closer to zero indicates less activity was subtracted from the EEG response during CCA artifact correction.

Acknowledgments

The authors would like to thank Maansi Desai, Mary Lowery, and Ian Griffith for their assistance in data collection. The authors additionally would like to thank Stéphanie Riès for her assistance with the preprocessing steps of the experiment and Alexander Etz for his statistical consultation during the peer review process.

Reprint requests should be sent to Liberty S. Hamilton, Department of Speech, Language, and Hearing Sciences, The University of Texas at Austin, 2504A Whitis Ave, Austin TX, 78712-1139, United States, or via e-mail: liberty.hamilton@austin.utexas.edu.

Data Availability Statement

Code for reproducing the analyses in this article can be found at https://github.com/HamiltonLabUT/speaker_induced_suppression_EEG/. The EEG data set and corresponding event files can be downloaded at https://doi.org/10.17605/OSF.IO/FNRD9.

Author Contributions

Garret L. Kurteff: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Rosemary A. Lester-Smith: Conceptualization; Investigation; Methodology; Writing—Review & editing. Amanda Martinez: Data curation; Investigation; Writing—Review & editing. Nicole Currens: Data curation; Investigation; Writing—Review & editing. Jade Holder: Data curation; Investigation; Writing—Review & editing. Cassandra Villarreal: Data curation; Investigation; Writing—Review & editing. Valerie R. Mercado: Data curation; Investigation; Writing—Review & editing. Christopher Truong: Data curation; Investigation; Writing—Review & editing. Claire Huber: Data curation; Investigation; Writing—Review & editing. Paranjaya Pokharel: Data curation; Investigation; Writing—Review & editing. Liberty S. Hamilton: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing.

Funding Information

Funding was provided from startup funds (to L. S. H.) from The University of Texas at Austin, Department of Speech, Language, and Hearing Science, and the Department of Neurology.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .492; W/M = .328; M/W = .082; W/W = .098.

REFERENCES

Appelbaum, I. (1996). The lack of invariance problem and the goal of speech perception. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* (Vol. 3, pp. 1541–1544). https://doi.org/10.1109/ICSLP.1996.607912

Astheimer, L. B., & Sanders, L. D. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia*, *49*, 3512–3516. https://doi.org/10.1016/j.neuropsychologia.2011.08.014, PubMed: 21875609

Barry, R. J., Kirkaikul, S., & Hodder, D. (2000). EEG alpha activity and the ERP to target stimuli in an auditory oddball paradigm. *International Journal of Psychophysiology*, *39*, 39–50. https://doi.org/10.1016/s0167-8760(00)00114-8, PubMed: 11120346

Behroozmand, R., & Larson, C. R. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neuroscience*, *12*, 54. https://doi.org/10.1186/1471-2202-12-54, PubMed: 21645406

Bendixen, A., Scharinger, M., Strauß, A., & Obleser, J. (2014). Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments. *Cortex*, *53*, 9–26. https://doi.org/10.1016/j.cortex.2014.01.001, PubMed: 24561233

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188. https://doi.org/10.1214/aos/1013699998

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer [Computer program]*. Version 5.53.51.

Bradley, A., Yao, J., Dewald, J., & Richter, C. (2016). Evaluation of electroencephalography source localization algorithms with multiple cortical sources. *PLoS One*, *11*, e0147266. https://doi.org/10.1371/journal.pone.0147266, PubMed: 26809000

Brumberg, J. S., & Pitt, K. M. (2019). Motor-induced suppression of the N100 event-related potential during motor imagery control of a speech synthesizer brain–computer interface. *Journal of Speech, Language, and Hearing Research*, *62*, 2133–2140. https://doi.org/10.1044/2019_JSLHR-S-MSC18-18-0198, PubMed: 31306609

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 629–647. https://doi.org/10.1002/wcs.63, PubMed: 23946864

Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, *98*, 1042–1054. https://doi.org/10.1016/j.neuron.2018.04.031, PubMed: 29779940

Chen, X., Xu, X., Liu, A., Lee, S., Chen, X., Zhang, X., et al. (2019). Removal of muscle artifacts from the EEG: A review and recommendations. *IEEE Sensors Journal*, *19*, 5353–5368. https://doi.org/10.1109/JSEN.2019.2906572

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife*, *5*, e12577. https://doi.org/10.7554/eLife.12577, PubMed: 26943778

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). London, UK: Routledge. https://doi.org/10.4324/9780203771587

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (MTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604, PubMed: 27965557

De Clercq, W., Vergult, A., Vanrumste, B., Van Paesschen, W., & Van Huffel, S. (2006). Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *IEEE Transactions on Bio-Medical Engineering*, *53*, 2583–2587. https://doi.org/10.1109/TBME.2006.879459, PubMed: 17153216

De Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*, 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017, PubMed: 28588065

de Vos, D. M., Riès, S. K., Vanderperren, K., Vanrumste, B., Alario, F. X., Huffel, V. S., et al. (2010). Removal of muscle artifacts from EEG recordings of spoken language production. *Neuroinformatics*, *8*, 135–150. https://doi.org/10.1007/s12021-010-9071-0, PubMed: 20480401

Desai, M., Holder, J., Villarreal, C., Clark, N., Hoang, B., & Hamilton, L. S. (2021). Generalizable EEG encoding models with naturalistic audiovisual stimuli. *Journal of Neuroscience*, *41*, 8946–8962. https://doi.org/10.1523/JNEUROSCI.2891-20.2021, PubMed: 34503996

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030, PubMed: 26412129

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., et al. (2015). Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences, U.S.A.*, *112*, 2871–2875. https://doi.org/10.1073/pnas.1414491112, PubMed: 25730850

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4, PubMed: 35260860

Gómez-Herrero, G. (2007). Automatic artifact removal (AAR) Toolbox v1. 3 (Release 09.12. 2007) for MATLAB. Tampere University of Technology. https://germangh.github.io/pubs/aardoc07.pdf

Gonzalez Castro, L. N., Hadjiosif, A. M., Hemphill, M. A., & Smith, M. A. (2014). Environmental consistency determines the rate of motor adaptation. *Current Biology*, *24*, 1050–1061. https://doi.org/10.1016/j.cub.2014.03.049, PubMed: 24794296

Goregliad Fjaellingsdal, T., Schwenke, D., Scherbaum, S., Kuhlen, A. K., Bögels, S., Meekes, J., et al. (2020). Expectancy effects in the EEG during joint and spontaneous word-by-word sentence production in German. *Scientific Reports*, *10*, 5460. https://doi.org/10.1038/s41598-020-62155-z, PubMed: 32214133

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027, PubMed: 24161808

Greenlee, J. D. W., Behroozmand, R., Larson, C. R., Jackson, A. W., Chen, F., Hansen, D. R., et al. (2013). Sensory-motor interactions for vocal pitch monitoring in non-primary human auditory cortex. *PLoS One*, *8*, e60783. https://doi.org/10.1371/journal.pone.0060783, PubMed: 23577157

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, *28*, 1860–1871. https://doi.org/10.1016/j.cub.2018.04.033, PubMed: 29861132

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*, 573–582. https://doi.org/10.1080/23273798.2018.1499946, PubMed: 32656294

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*, 4626–4639. https://doi.org/10.1016/j.cell.2021.07.019, PubMed: 34411517

Hashimoto, Y., & Sakai, K. L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. *Human Brain Mapping*, *20*, 22–28. https://doi.org/10.1002/hbm.10119, PubMed: 12953303

Hawco, C. S., Jones, J. A., Ferretti, T. R., & Keough, D. (2009). ERP correlates of online monitoring of auditory feedback during vocalization. *Psychophysiology*, *46*, 1216–1225. https://doi.org/10.1111/j.1469-8986.2009.00875.x, PubMed: 19674393

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35. https://doi.org/10.3389/fnsys.2014.00035, PubMed: 24672438

Heinks-Maldonado, T. H., Mathalon, D. H., Houde, J. F., Gray, M., Faustman, W. O., & Ford, J. M. (2007). Relationship of imprecise corollary discharge in schizophrenia to auditory hallucinations. *Archives of General Psychiatry*, *64*, 286–296. https://doi.org/10.1001/archpsyc.64.3.286, PubMed: 17339517

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *NeuroReport*, *17*, 1375–1379. https://doi.org/10.1097/01.wnr.0000233102.43526.e9, PubMed: 16932142

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., et al. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, *9*, 217. https://doi.org/10.3389/fnins.2015.00217, PubMed: 26124702

Houde, J. F., & Chang, E. F. (2015). The cortical computations underlying feedback control in vocal production. *Current Opinion in Neurobiology*, *33*, 174–181. https://doi.org/10.1016/j.conb.2015.04.006, PubMed: 25989242

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*, 82. https://doi.org/10.3389/fnhum.2011.00082, PubMed: 22046152

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*, *14*, 1125–1138. https://doi.org/10.1162/089892902760807140, PubMed: 12495520

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453–458. https://doi.org/10.1038/nature17637, PubMed: 27121839

Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). Probing artificial neural networks: Insights from neuroscience. *arXiv:2104.08197*. https://doi.org/10.48550/arXiv.2104.08197

Jahanshahi, M., & Hallett, M. (2003). *The Bereitschaftspotential: Movement-related cortical potentials*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4615-0189-3

Jiang, X., Bian, G., & Tian, Z. (2019). Removal of artifacts from EEG signals: A review. *Sensors*, *19*, 987. https://doi.org/10.3390/s19050987, PubMed: 30813520

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America*, *108*, 1246–1251. https://doi.org/10.1121/1.1288414, PubMed: 11008824

Kearney, E., & Guenther, F. H. (2019). Articulating: The neural mechanisms of speech production. *Language, Cognition and Neuroscience*, *34*, 1214–1229. https://doi.org/10.1080/23273798.2019.1589541, PubMed: 31777753

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. https://doi.org/10.2307/2533558, PubMed: 9333350

Khalighinejad, B., da Silva, G. C., & Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, *37*, 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017, PubMed: 28119400

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lester-Smith, R. A., Daliri, A., Enos, N., Abur, D., Lupiani, A. A., Letcher, S., et al. (2020). The relation of articulatory and vocal auditory–motor control in typical speakers. *Journal of Speech, Language, and Hearing Research*, *63*, 3628–3642. https://doi.org/10.1044/2020_jslhr-20-00192, PubMed: 33079610

Levelt, W. J. M. (1993). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Lightfoot, G. (2016). Summary of the N1-P2 cortical auditory evoked potential to estimate the auditory threshold in adults. *Seminars in Hearing*, *37*, 1–8. https://doi.org/10.1055/s-0035-1570334, PubMed: 27587918

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press. https://mitpress.mit.edu/9780262525855/an-introduction-to-the-event-related-potential-technique/

Martikainen, M. H., Kaneko, K., & Hari, R. (2005). Suppressed responses to self-triggered sounds in the human auditory cortex. *Cerebral Cortex*, *15*, 299–302. https://doi.org/10.1093/cercor/bhh131, PubMed: 15238430

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2019). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*, *31*, 327–338. https://doi.org/10.1162/jocn_e_01276, PubMed: 29916793

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*, 1006–1010. https://doi.org/10.1126/science.1245994, PubMed: 24482117

Mollaei, F., Shiller, D. M., Baum, S. R., & Gracco, V. L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Research*, *1646*, 269–277. https://doi.org/10.1016/j.brainres.2016.06.013, PubMed: 27288701

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*, 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026, PubMed: 17931964

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience*, *33*, 16110–16116. https://doi.org/10.1523/JNEUROSCI.2137-13.2013, PubMed: 24107944

Okada, K., Matchin, W., & Hickok, G. (2018). Phonological feature repetition suppression in the left inferior frontal gyrus. *Journal of Cognitive Neuroscience*, *30*, 1549–1557. https://doi.org/10.1162/jocn_a_01287, PubMed: 29877763

Ozker, M., Doyle, W., Devinsky, O., & Flinker, A. (2022). A cortical network processes auditory error signals during human speech production to maintain fluency. *PLoS Biology*, *20*, e3001493. https://doi.org/10.1371/journal.pbio.3001493, PubMed: 35113857

Parrell, B., Ramanarayanan, V., Nagarajan, S. S., & Houde, J. F. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLoS Computational Biology*, *15*, e1007321. https://doi.org/10.1371/journal.pcbi.1007321, PubMed: 31479444

Perkell, J., Matthies, M., Lane, H., Guenther, F. H., Wilhelms-Tricarico, R., Wozniak, J., et al. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication*, *22*, 227–250. https://doi.org/10.1016/S0167-6393(97)00026-5

Poeppel, D., & Monahan, P. J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes*, *26*, 935–951. https://doi.org/10.1080/01690965.2010.493301

Rastatter, M., & De Jarnette, G. (1984). EMG activity with the jaw fixed of orbicularis Oris superior, orbicularis oris inferior and masseter muscles of articulatory disordered children. *Perceptual and Motor Skills*, *58*, 286. https://doi.org/10.2466/pms.1984.58.1.286, PubMed: 6718194

Riès, S. K., Janssen, N., Burle, B., & Alario, F. X. (2013). Response-locked brain dynamics of word production. *PLoS One*, *8*, e58197. https://doi.org/10.1371/journal.pone.0058197, PubMed: 23554876

Riès, S. K., Pinet, S., Nozari, N. B., & Knight, R. T. (2021). Characterizing multi-word speech production using event-related potentials. *Psychophysiology*, *58*, e13788. https://doi.org/10.1111/psyp.13788, PubMed: 33569829

Schneider, D. M., Nelson, A., & Mooney, R. (2014). A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature*, *513*, 189–194. https://doi.org/10.1038/nature13724, PubMed: 25162524

Shuster, L. I. (2003). fMRI and normal speech production. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, *13*, 16–21. https://doi.org/10.1044/nnsld13.3.16

Singh, T., Phillip, L., Behroozmand, R., Gleichgerrcht, E., Piai, V., Fridriksson, J., et al. (2018). Pre-articulatory electrical activity associated with correct naming in individuals with aphasia. *Brain and Language*, *177–178*, 1–6. https://doi.org/10.1016/j.bandl.2018.01.002, PubMed: 29421267

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, *164*, 77–105. https://doi.org/10.1016/j.bandl.2016.10.004, PubMed: 27821280

Stepp, C. E. (2012). Surface electromyography for speech and swallowing systems: Measurement, analysis, and interpretation. *Journal of Speech, Language, and Hearing Research*, *55*, 1232–1246. https://doi.org/10.1044/1092-4388(2011/11-0214, PubMed: 22232412

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, *20*, 2315–2331. https://doi.org/10.1523/JNEUROSCI.20-06-02315.2000, PubMed: 10704507

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*, 952–981. https://doi.org/10.1080/01690960903498424, PubMed: 23667281

Turin, G. (1960). An introduction to matched filters. *IRE Transactions on Information Theory*, *6*, 311–329. https://doi.org/10.1109/TIT.1960.1057571

Van Eijden, T. M., Blanksma, N. G., & Brugman, P. (1993). Amplitude and timing of EMG activity in the human masseter muscle during selected motor tasks. *Journal of Dental Research*, *72*, 599–606. https://doi.org/10.1177/00220345930720030801, PubMed: 8450119

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*, 989–994. https://doi.org/10.1016/s0028-3932(02)00316-0, PubMed: 12667534

Wohlert, A. B. (1993). Event-related brain potentials preceding speech and nonspeech oral movements of varying complexity. *Journal of Speech and Hearing Research*, *36*, 897–905. https://doi.org/10.1044/jshr.3605.897, PubMed: 8246478

Wrench, A. (1999). The MOCHA-TIMIT articulatory database. https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

Yoshida, K., Kaji, R., Hamano, T., Kohara, N., Kimura, J., & Iizuka, T. (1999). Cortical distribution of Bereitschaftspotential and negative slope potential preceding mouth-opening movements in humans. *Archives of Oral Biology*, *44*, 183–190. https://doi.org/10.1016/s0003-9969(98)00122-8, PubMed: 10206336

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*, 3878. https://doi.org/10.1121/1.2935783

Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 992–96). https://doi.org/10.1109/ICASSP.2015.7178118

Zheng, Z. Z., Munhall, K. G., & Johnsrude, I. S. (2010). Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. *Journal of Cognitive Neuroscience*, *22*, 1770–1781. https://doi.org/10.1162/jocn.2009.21324, PubMed: 19642886