Cortical Representation of the Glottal Events during Speech Production

John P. Veillette[1†]
Jacob Rosen[1]
Daniel Margoliash[1,2]
Howard C. Nusbaum[1]

[1]Department of Psychology, University of Chicago
[2]Department of Organismal Biology and Anatomy, University of Chicago

[†]Correspondence should be addressed to John P. Veillette; E-mail: johnv@uchicago.edu

Author Contributions (CRediT Taxonomy)

D.M.: Conceptualization, Funding acquisition, and Writing - review & editing. **H.C.N.:** Conceptualization, Funding acquisition, Resources, Supervision, and Writing - review & editing. **J.P.V.:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing. **J.R.:** Data curation, Investigation, and Writing - review & editing.

Acknowledgements

This work was supported by a National Science Foundation grant (NSF BCS 1835181) to D.M and H.C.N.; J.P.V. was supported by an NSF GRFP (NSF DGE 1746045) and a Neubauer Family Distinguished Doctoral Fellowship.

Competing Interests

The authors declare no competing interests.

Abstract

To produce complex motor behaviors such as speech, the nervous system must accommodate the fact that sensory feedback is delayed relative to actual movements; by most accounts, this is accomplished using an internal model that predicts the current state of the body's periphery from recent motor system output. Here we show that onsets of events in the human glottal waveform, measured via electroglottography, are encoded in the human electroencephalogram (EEG) during speech production, maximally reflected at zero time lag. Conversely, we show that glottal event times can be decoded from the EEG. Furthermore, after prolonged exposure to delayed auditory feedback, subjects show a robust recalibration of their behaviorally observed threshold for detecting auditory-motor mismatches, and decoding models that perform best in normal speaking conditions also showed a shift in the predicted times of glottal events relative to actual events. This suggests that decoding performance is driven by plastic internal representations of peripheral event timing (while ruling out movement artifact concerns). Our results provide a missing component for a mechanism to associate specific feedback events with the neurons that gave rise to those movements, mirroring the observation of synchronous bursting of songbird HVC neurons with events in the trajectories of the biophysical parameters that control bird voicing.

Main

Speech production requires coordination of over a hundred muscles to achieve desired acoustic output (Simonyan & Horwitz, 2011). Were auditory feedback immediate, acoustic outcomes could be assigned to articulatory movements in real-time, allowing control policies to be learned in a "model-free" manner as in contemporary policy-based reinforcement learning systems (Kakade, 2001). However, sensory transduction and multiple stages of neuronal transmission all result in intrinsic delays of the arrival of feedback to forebrain structures often associated with learned movements (Wolpert et al., 2011; Wolpert & Ghahramani, 2000). Consequently, vocal learning – and motor learning in general – is thought to require a "forward" model to actively predict current and future peripheral states and, in turn, the resulting sensory feedback from the recent motor output; discrepancies between predicted and actual feedback are used to update the forward model, which can then be leveraged offline to adjust an "inverse" model (i.e. control policy) which proposes movements to achieve desired states (Tourville & Guenther, 2011). This computational account has become a guiding framework for understanding both normal motor behavior and the symptomology of speech motor deficits such as stuttering (Bradshaw et al., 2021). But while forward model updates have been proposed to occur through reinforcement learning mechanisms (Sutton & Barto, 2018), the task of assigning outcomes to actions when feedback is temporally delayed – the temporal/hindsight credit assignment problem – remains a standing challenge in theoretical and applied reinforcement learning (Harutyunyan et al., 2019).

A plausible mechanism has begun to emerge from research on birdsong production from zebra finches. In the HVC (premotor/association cortex analog), certain neurons produce spike bursts and interneurons are suppressed within milliseconds of transitions between the syringeal pressure and tension gestures used to control birds' analog to the human glottis within their syrinx (Amador et al., 2013). As bursts occur too late to have triggered movements but too early to represent sensory feedback, this synchronous activity has been interpreted as a state prediction. More recent work has shown that some primary motor cortex analog projecting HVC neurons also burst at syllable onsets or offsets, although the timing of these burst relative to syringeal gestures has yet to be determined (Moll et al., 2023). Notably, pressure-tension gesture transitions are quickly reflected in the derivative of the song envelope (Boari et al., 2015), and some Field L (auditory cortex analog) neurons selectively respond to such derivative information in auditory stimuli (Sharpee et al., 2011). As some HVC neurons are prone to repeat bursts after fixed intervals following their initial recruitment during song (Daou & Margoliash, 2020; Fetterman & Margoliash, 2023), time-delayed bursts representing gestures and corresponding sensory feedback events should be simultaneously available for associative learning (Joris et al., 1998).

A notable feature of this account, should it generalize to humans, is that temporal coherence between gesture transitions and resulting acoustic events eliminates the need to predict movements and their specific auditory consequences sequentially as in extant speech production models (Golfinopoulos et al., 2010; Guenther et al., 2006), potentially explaining the perplexing finding that playing tones that are amplitude-modulated to match the speech envelope with a delay has similar effects on speech production as delayed auditory feedback (DAF) (Howell & Archer, 1984). Consistent with this theory, auditory responses in human superior temporal gyrus prominently reflect acoustic event onsets (Oganian & Chang, 2019). However, cortical tracking of glottal events has yet to be demonstrated. We recorded subjects' electroencephalography (EEG),

audio, and glottis vibrations using electroglottograph (EGG) during speech production. We fit encoding models to test whether glottis event onsets explain EEG variance beyond previously identified features. Further, we exposed subjects to a period of DAF to, in principle, temporally shift their forward model predictions, reflected behaviorally in their threshold for detecting auditory feedback delays (Yamamoto & Kawabata, 2014). Finally, we tested whether a model trained to decode glottal events from brain activity prior to prolonged DAF exposure makes temporally shifted predictions after this threshold shift, reflecting subjects' own updated predictions.

Task Design

Throughout the experiment, subjects heard their auditory feedback through insert earphones while they spoke into a microphone. Subjects read sentences that are presented on a monitor in front of them during (1) a **baseline block** with no delay, i.e. normal speaking conditions, and (2) a **pretest block** in which we imposed a random delay between 0 and 250 ms in their auditory feedback on each trial. To assess whether glottal event times were encoded in the EEG, we trained a multivariate encoding model that predicted the continuous EEG from the speech audio envelope, speech audio onsets, the EGG envelope, and the EGG onsets, and we compared its predictive performance to a model that that excluded EGG audio onsets. Encoding models were trained during the pretest block, where audio and glottal events were temporally decoupled by design, and tested during the baseline block to assess generalization performance in normal speaking conditions.

Subjects then completed (3) an **exposure block** with a constant 200 ms delay, and (4) a **posttest block** that was identical to the pretest block. During the pretest and posttest blocks, subjects were asked to report, on each trial, whether they detected a delay in their auditory feedback; a multilevel logistic model was fit to find the threshold at which they could detect a delay 50% of the time in both blocks, so we could assess perceptual shift as a result of DAF exposure. Then, we trained a decoding model to predict glottal event onsets on the pretest block and tested it on the posttest block (and vice versa) to assess whether the predictions of models that achieve high decoding accuracy are systematically affected by exposure to delayed auditory feedback, consistent with decoding performance being driven by an internal representation that is plastic given sensorimotor experience.

Event onsets in the glottal waveform are reflected in the EEG during speech production.

The encoding model that incorporates all features – speech audio envelope, audio onsets, EGG envelope, and EGG onsets – performed robustly above chance (TFCE statistic = 11.19, $p = 0.0001$), as did the model that excludes EGG onsets (TFCE = 9.95, $p = 0.0001$). However, including the EGG onsets in the model did improve out-of-sample prediction of the EEG signal (TFCE = 2.38, $p = 0.0085$), indicating that the time of events in the EGG explains additional variance in cortical neural activity that is not explained by these other features.

The EEG temporal response function (TRF) to EGG onsets, visualized in Figure 1, peaks roughly synchronously to the actual events in the glottal waveform. This is the theoretically predicted timescale of a forward model's predictive representations, evidence of which is observed

both in the motivating songbird studies and, though outside of the auditory system, in primates (Amador et al., 2013; Mulliken et al., 2008). However, a synchronous TRF would also be predicted by a movement artifact; our subsequent decoding analyses were designed to assess the possibility of such a confound.

Glottal event onsets can be decoded from the EEG.

All four features could be decoded from the EEG with above-chance cross-validation performance, with all familywise error rate (FWER)-corrected $p < 0.0001$. The out-of-sample correlation [95% CI] between decoded and actual feature time series was $r = 0.233$ [0.199, 0.268])for the audio envelope, r = 0.101 [0.086, 0.117] for audio onsets, $r = 0.349$ [0.312, 0.382] for EGG envelope, and $r = 0.136$ [0.120, 0.152] for EGG onsets. Subject-level decoding performances are shown in Figure 2a.

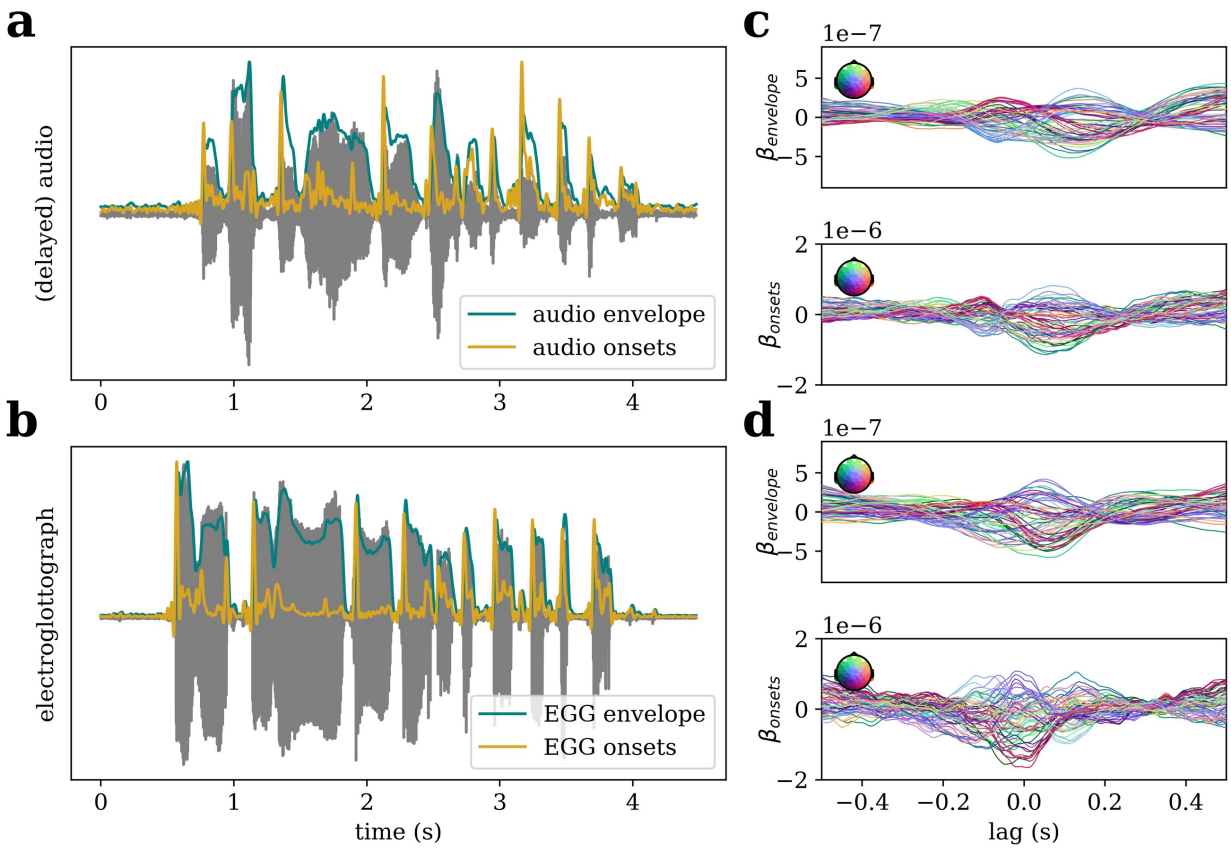

Figure 1: EEG temporal response functions to features of glottal waveform and auditory feedback waveform during speech production. From (a) the speech audio, played back to the subject at a delay subsequent to the baseline block, and from (b) the glottal waveform, we compute an amplitude envelope that approximates the frequency response of the cochlea and an "onset" time series that reflects the positive first derivative of the amplitude across all frequency bands – thus the onset of spectral events. Temporal response functions (TRFs), the EEG responses to these features of the (c) auditory feedback or (d) glottal waveform, are shown; a large coefficient in the TRF at, for example, lag 0.1 s means that a change in the value of that feature manifests in the EEG 0.1 seconds after that change.

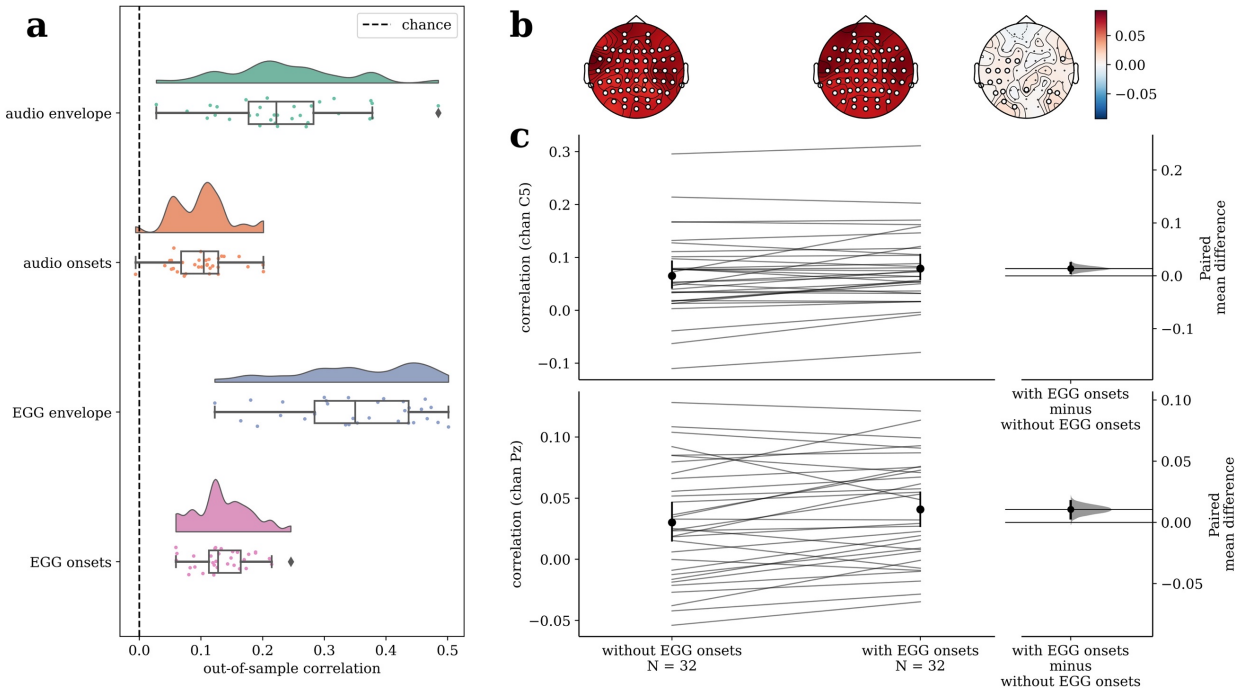

Figure 2: Out-of-sample predictive performance of decoding and encoding models. (a) The per-subject correlations between actual time series of the audio/EGG features and those decoded from the EEG. (b) The cross-validated correlation between actual EEG and that predicted from the audio/EGG features, shown for all electrodes. The model on the left excludes the EGG onsets as a predictor, the middle includes EGG onsets, and the difference in performance is shown on the right. White circles mark electrodes that are significant after multiple comparisons correction. (c) The same encoding model performances are shown for a central and a parietal electrode; paired correlations for each subject, their bootstrap means with 95% confidence intervals, and the bootstrap distribution of the mean differences are shown. All decoding and encoding models visualized here were trained in the pretest block (i.e. before exposure) and tested in the baseline block – in other words, in normal speaking conditions.

Perception of auditory-motor synchrony temporally recalibrates after prolonged exposure to delayed auditory feedback.

The group-level parameters [95% highest posterior density interval (HPDI)] of the logistic multilevel model (time units in milliseconds) used to measure delay detection threshold were $\beta_{intercept}$ = -2.19 [-2.64, -1.70], β_{delay} = 36.6 [31.16, 42.25], $\beta_{adaption}$ = -0.97 [-1.52, -0.414], $\beta_{delay \times adaption}$ = 2.61 [-3.18, 8.47]. The detection threshold, or the delay at which the probability of detection is 50%, shifted from mean 59.8 ms pre-exposure to 80.8 ms post-exposure; in other words, subjects' detection thresholds shifted by about 20.9 ms [10.7, 31.6] after the exposure block. Trial-level behavioral data, group-level logistic curves, and both group- and subject-level estimates of threshold shift are visualized in Figure 3.

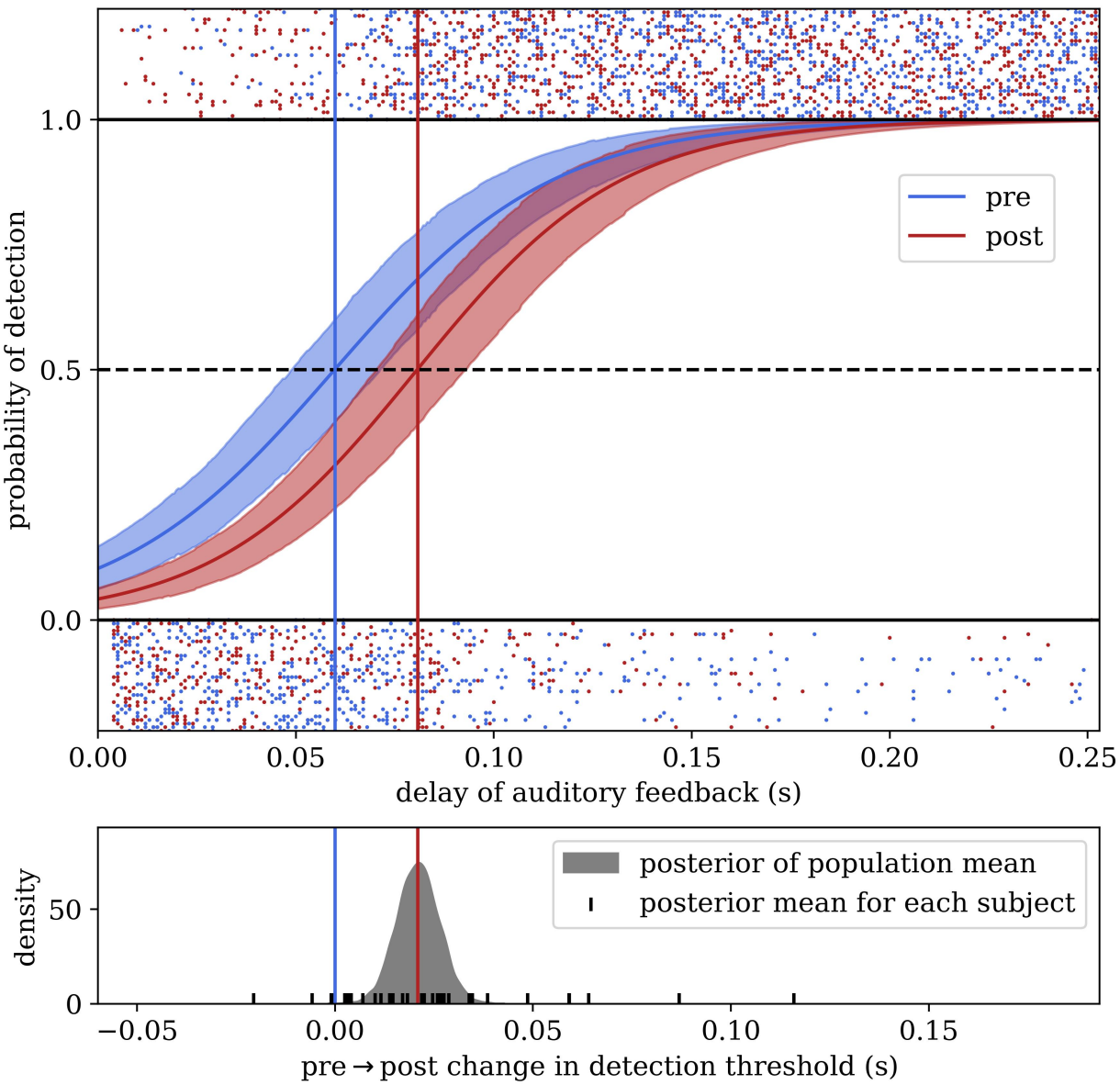

Figure 3: Behavioral shift in delay detection threshold following exposure to delayed auditory feedback. (top) The group-level logistic regression functions describing the probability of detecting an auditory-motor asynchrony in the pretest and posttest blocks, which flank an exposure block where subjects are exposed to a constant auditory feedback delay. Trial-by-trial responses are shown above and below the logistic curves, with rows corresponding to subjects. (bottom) The posterior distribution of the population/group average change in the delay detection threshold, with point estimates shown for each individual subject.

Decoded glottal event times are affected by perceptual recalibration.

When we trained decoding models to predict EGG onsets from brain activity on the pretest block, we found that models which performed better in normal speaking conditions (i.e. baseline block) predicted later onset times relative to the actual EGG events when tested during posttest (Spearman's rank correlation = 0.410, p = 0.0198). In alignment with our *a priori* hypothesis, high performing (i.e. above median) decoders predicted event onsets that, on average, were slightly delayed relative to actual events (lag = 0.003 s [0.000, 0.005]; t(31) = 1.86, p = 0.0413), mirroring

the perceptual temporal recalibration observed in subjects' behavior. In contrast, the 95% confidence interval for lower performing decoders overlaps with zero delay (lag = -0.003 s [-0.008, 0.001]).

Models trained on posttest and tested at pretest showed the opposite relationship (Spearman's rank correlation = -0.350, p = 0.0497). High performing decoders predicted EGG onset times that were earlier than actual event onsets (-0.003 s [-0.005, 0.000]; t(31) = -1.86, p = 0.0413), but again the confidence interval for low-performing decoders contains zero lag (-0.002 s [-0.003, 0.007]). The fact that decoder performance is monotonically related to how much a decoder is affected by exposure to an auditory-motor delay rules out the possibility that decoding performance is explained by a movement artifact or non-plastic sensory response, which would be invariant to temporal recalibration. (We do note that, despite this group trend, lags of zero are still common in the sample, so it is possible that such confounds still contribute to decoding performance – they just cannot fully explain it.)

The pre-to-posttest lags and post-to-pretest lags were not significantly different from one another for low correlation decoders (-0.005 s [-0.013, 0.002]; paired permutation test p = 0.177), as would have been expected for real (non-noise) shifts, but they were different for high performing decoders (0.005 s [0.0004, 0.011], paired permutation test p = 0.046). The difference-in-difference (i.e. interaction) is significant (0.012 [0.001, 0.020]; paired permutation p = 0.049), suggesting that the correlation between the lag of predictions and decoding performance is indeed accounted for by increased signal-to-noise in high performing decoding models. Prediction lags and their differences are visualized in Figure 4.

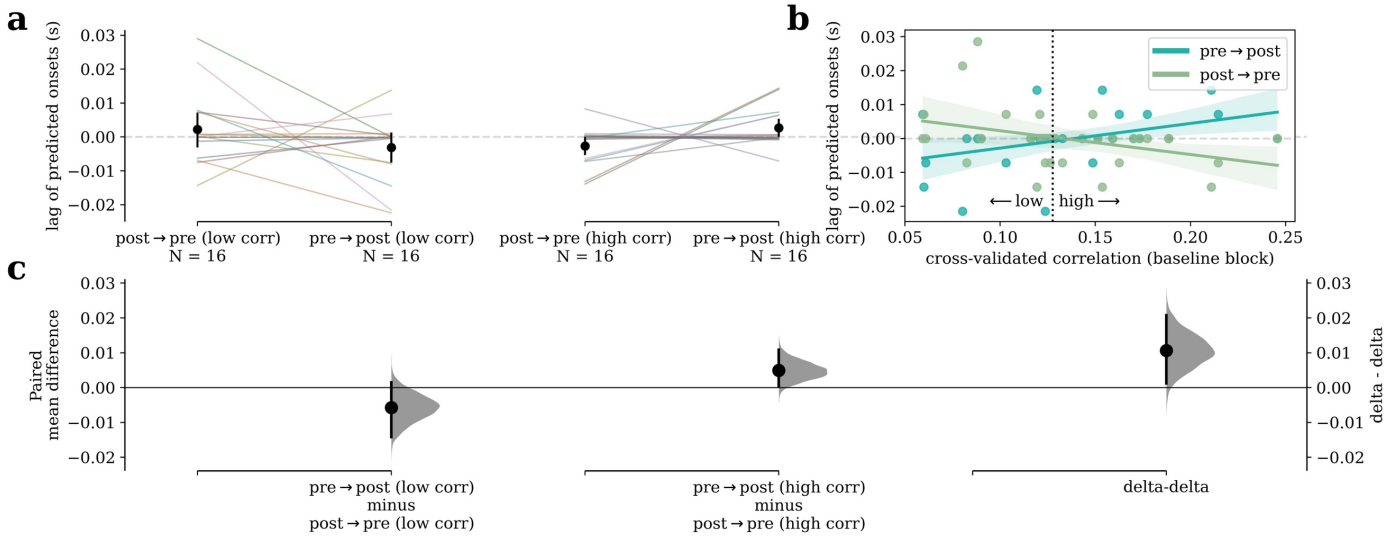

Figure 4: Temporal shifts in electroglottograph event times decoded from EEG with exposure to delayed auditory feedback. (a) The lags, with bootstrap 95% confidence intervals, of predicted EGG event onsets relative to actual onsets for decoding models trained pre-exposure and tested post-exposure, as well as for models trained in post-exposure and tested pre-exposure. Models are median-split by their decoding performance in normal speaking conditions, as illustrated in (b) where all lags are shown with their best linear fits and bootstrap 95% confidence bands. (c) Bootstrap distributions of paired differences between pre-to-post and post-to-pretest prediction lags for low and high performing decoders, as well as for their difference-in-differences, which indicates that exposure-contingent temporal shifts in decoder predictions are larger for high performing models.

Parallels to birdsong suggest a common mechanism of vocal learning across species.

The present study presents evidence that, in parallel to observations in singing birds (Amador et al., 2013), neural activity in human cortex tracks the timing of events in the glottal waveform during speech production in roughly temporally coordinate fashion. Acoustic events in the glottal waveform have a lawful temporal correspondence to the subglottal pressure and laryngeal tension gestures that are used to control the vibration of the glottis (Titze, 1988); the same is true for the equivalent avian structures (Boari et al., 2015; Perl et al., 2011). Moreover, the glottal waveform serves as the acoustic source that, once filtered through the upper vocal tract, results in the speech waveform (Titze, 1998). Thus, under normal speaking conditions, event onsets in the glottal waveform have a one-to-one correspondence with subsequent acoustic events in the auditory stream, and such onset events are known to be prominently reflected in the auditory cortical representation of speech (Oganian & Chang, 2019). In other words, cortical tracking of glottal gestural/acoustic event times could contribute to the circuit resolving the temporal credit assignment problem – based on temporal coherence between event times – in the control of voicing (Harutyunyan et al., 2019). Work in birdsong helps to elaborate these observations. Synchronous activity during singing between zebra finch HVC neurons and syringeal gestures (Amador et al., 2013) has been proposed to be a manifestation of the anticipated synchronization characteristic of certain coupled dynamical systems, which yield zero lag relationships between hierarchically organized structures (Alonso et al., 2015; Dima et al., 2018). The patterns of respiratory activity in singing canaries are also predicted by such models (Alonso et al., 2015; Dima et al., 2018), and such modeling emphasizes the importance of ascending input from brainstem respiratory centers (Ashmore et al., 2005, 2008; Gibb et al., 2009). These observations provide elements for simultaneous representations of gestures and their acoustic outcomes, which would be linked by well-documented coincidence detection mechanisms (Joris et al., 1998; Matell & Meck, 2004).

It is worth elaborating that such a mechanism, emphasizing event onsets rather than event content as described, would not replace the content-specific forward model predictions posited by extant models of speech motor control (Tourville & Guenther, 2011); indeed, frequency-specific suppression of sounds resulting from movement is well established in mammalian auditory cortex (Schneider et al., 2018; Schneider & Mooney, 2018). Rather, representing articulatory movements at a time delay allows gestures to be matched to acoustic events with temporally coherent onsets, which affords learning content-specific action-outcome mappings via simple associative mechanisms. An important enabling condition for such a parsimonious system is temporal correspondence between articulatory and acoustic events at the sound source, which is a consequence of the nonlinear dynamics of the vocal folds – in which continuous changes in pressure/tension control parameters can result in discontinuous transitions in acoustic output, translating articulatory gestures into discrete event onsets (Boari et al., 2015; Sitt et al., 2008). In this view, then, the learning of a forward model for the control of voicing is emergently facilitated by the manner in which the central nervous system interacts with the vocal organ, circumventing the need for specialized CNS mechanisms to infer actions-outcome pairs post hoc.

Methods

1. Ethics and Recruitment

We recruited 32 university students (ages 18-22; 15 males, 17 females) over our department's study recruitment system. Subjects self-reported being right handed, though 3 were considered ambidextrous by the Edinburgh Handedness Inventory (Oldfield, 2013). Informed consent was obtained prior to the experiment, and procedures were approved by the Social and Behavioral Sciences Institutional Review Board at the University of Chicago (IRB21-1402).

2. Experimental Design

After applying EEG and EGG electrodes to the subjects' heads and necks, respectively, they sat in front of a monitor. In each trial of the subsequent task, subjects were asked to read one of the phonetically balanced Harvard Sentences, which was displayed on the monitor ("IEEE Recommended Practice for Speech Quality Measurements," 1969). We controlled the trial-by-trial latency of auditory feedback, recorded through a microphone (Behringer ECM-8000) and played back through insert earphones (Etymotic ER-3C) which blocked out external sound, using a custom Python tool. Since such a system requires digitizing audio and then converting back to analog to play back to the subject, there is an intrinsic minimum hardware/software delay – that is, the actual delay between input and output audio when we specify a delay of zero – which we empirically measured as 4 milliseconds by recording input and output audio simultaneously on our EEG amplifier. For comparison, the intrinsic latency of other auditory feedback manipulation systems used in the literature tends to exceed 15 milliseconds (Kim et al., 2020). The delays recorded in our open dataset (and used in our analysis) have been corrected to account for this minimum delay.

The task consisted of four blocks: (1) a baseline block in which subjects read 20 sentences with no imposed feedback delay, (2) a pretest block in which subjects read 50 sentences each with a random feedback delay, uniformly distributed between 0 and 0.25 seconds, (3) an exposure block in which subjects read 80 sentences with a constant 200 ms delay, and (4) a posttest block, identical to the pretest block. After each trial in pretest and posttest blocks, subjects were asked to indicate on a keyboard if they detected a delay between their movements and the sound of their voice. Sentence prompts were unique for each subject.

3. Statistical Analysis

From both the delayed speech audio and the EGG, we computed spectrograms between 30 and 5000 Hz using a gammatone filter bank, which approximates the frequency response of the human cochlea and has been noted to outperform other sound frequency decompositions when predicting EEG from the speech envelope (Issa et al., 2024). The envelope of both the audio and EGG was computed by averaging the amplitude across equivalent rectangular bandwidths (i.e. an approximation to the frequency bandwidths in human hearing). As in prior EEG and electrocorticography work, an acoustic event onset time series was computed by taking the positive first derivative within each equivalent rectangular bandwidths and then averaging across them (Brodbeck et al., 2023; Oganian & Chang, 2019). Envelopes (but not onsets) were log-

transformed, which better reflects the psychophysical curve for loudness and has been shown to improve prediction of EEG (Brodbeck et al., 2023). Subsequently, all features and the EEG were scaled to be of roughly equal variance using the *RobustScaler* in the *scikit-learn* package (Pedregosa et al., 2011). Feature scaling was performed within blocks, rather than across the whole dataset, to prevent train-test leakage during cross-validation.

For EEG encoding models, we estimated Temporal Response Functions as the coefficients of a linear ridge regression that predicts each electrode from values of the time-lagged audio and EGG envelopes from 0.5 seconds before to 0.5 seconds after each EEG time point using the *MNE-Python* package (Crosse et al., 2016; Gramfort et al., 2014). While it has recently become popular in the EEG literature (Brodbeck et al., 2023), this approach has a longer history in single-cell physiology (Theunissen et al., 2001) and is also frequently used, though referred to as "voxelwise-encoding," in the fMRI literature (Dupré la Tour et al., 2022; Huth et al., 2016). Encoding models were trained on the pretest block, where the randomly varied latency of the auditory feedback allowed the model to differentiate between the EGG and audio waveforms, which would otherwise be roughly synchronous, and then tested (i.e. cross-validated) under normal speaking conditions in the baseline block, where we computed the correlation between the predicted and actual values at each electrode to measure the encoding models' generalization performance. We fit both an encoding model that includes all four audio and EGG features and one excluding EGG onsets to test whether EGG event onsets explain variation in brain activity beyond that already explained by the audio envelope, audio onsets, and the EGG envelope. Cross-validated correlations of both models were statistically compared to chance (zero correlation) and to each other using threshold-free cluster enhancement to control the familywise error rate (Smith & Nichols, 2009).

Decoding models were fit using the same approach but in the reverse direction, predicting the values of each audio and EGG features from the -0.5 to 0.5 second time-lagged EEG electrodes. While only the decoding of the EGG onsets is of interest in the present study, we report the cross-validated decoding performance – again, trained on pretest and tested on baseline – for all features, with p-values comparing each to chance (zero correlation) computed using a permutation test and familywise error-rate corrected with the t-max procedure (Nichols & Holmes, 2002).

To manipulate the content of subjects' forward models, we exposed subjects to 0.2 second delayed auditory feedback throughout the exposure block, after which the same procedure as the pretest block was repeated in posttest. We estimated the shift in subjects' threshold for detecting auditory-motor asynchronies, where "threshold" is defined as where the probability of detection is 50%, using a Bayesian multilevel logistic regression estimated with the *PyMC* package (Patil et al., 2010). Under the assumption that subjects detect a delay when their auditory feedback is sufficiently different from the feedback they predict, this threshold is a behavioral index of the content of their forward model and, as a corollary, a pretest to posttest change in threshold indicates a change in their forward model.

If the EEG signal that reflects EGG event onsets is indeed an index of predictive cortical tracking of glottal events – in analog to HVC activity in songbirds (Amador et al., 2013) – then that signal should change from pretest to posttest as do subjects' detection thresholds. To test this hypothesis, we take those models which have been trained to predict EGG event onsets in the pretest block, and we generate its predicted onset time series for the posttest block; we then

compute the time lag at which the cross-correlation between the predicted and actual event onsets is maximal as an estimate of how much predictions have shifted in time between pretest and posttest. Additionally, we train a decoding model on the posttest block and compute the same cross-correlation lag testing on the pretest block.

These measurements allow us to assess the possibility that the signal driving decoding is (a) a response to the glottal waveform via another sensory modality, e.g. somatosensory, or through bone conducted auditory feedback, or it reflects (b) a movement/muscle artifact or contamination from the EGG recording itself. Should either of these possibilities account for above-chance decoding performance, then there should be no trend in pre-to-posttest prediction lags or post-to-pretest lags with increasing baseline performance; at-chance models will always average zero lag, since any peak in the cross-correlation function would be spurious, and higher performing models will tend toward zero lag as signal-to-noise for decoding from the confound signal, which should not be affected by perceptual recalibration following DAF, improves. Conversely, if the signal that drives decoding performance is related to subjects' forward model, then the predictions of decoders with higher signal-to-noise should be more affected. Thus, the alternative hypothesis predicts a monotonic trend in pre-to-posttest lags and an opposite trend in post-to-pretest lags as baseline decoding performance improves. Thus, we test the null hypothesis using a Spearman rank correlation between the cross-validated correlation (trained on pretest, tested on baseline) for each subject and their pre-to-posttest shift in decoded onsets.

It is not necessarily the case that a shift in detection threshold corresponds directly to a same-direction shift in the expected time of glottal events as predicted by subjects' forward models; psychophysical evidence indicates that humans also maintain a prior on the latency between action and sensory outcomes, and whether the expected movement time or expected action-outcome latency is updated following a prediction error would depend on the relative prior precision of the two estimates (Legaspi & Toyoizumi, 2019). However, the former explanation does make the distinct prediction that that pretest-trained decoder predictions will be delayed in the posttest block, and posttest-trained decoder predictions will be early in the pretest block – and again, these lags should be larger in models with higher decoding performance. Thus, we report two *a priori* one-tailed *t*-tests for the pre-to-posttest and the post-to-pretest prediction lags, comparing both to zero, in those subjects whose decoders perform best (i.e. above median) in normal speaking conditions. We additionally report estimation statistics (bootstrap confidence intervals) for the pre-to-posttest shifts, post-to-pretest shifts, and their paired differences, for high-performing and low-performing models separately (median split). We use the *DABEST-Python* package to compute and visualize bootstrap confidence intervals, which also generates two-tailed *p*-values for the pairwise differences in pre-to-post and post-to-pretest lags – which should be opposite in sign from one another if shifts due to DAF exposure are genuine – via permutation test, as well as for the difference-in-differences across the median split, which reflects whether exposure-contingent shifts increase (or decrease) with signal-to-noise; these too are reported for completeness.

4. EEG Acquisition and Preprocessing

EEG was acquired at 10,000 Hz using an actiCHamp amplifier (Brain Products, GmbH) with 60 active EEG electrodes, two electrooculogram (EOG) electrodes below the eyes, and two electromyogram (EMG) channels over the jaw muscles in front of the ears, referenced to their ipsilateral mastoid. The precise positions of each EEG electrode were measured and recorded using a CapTrak (Brain Products, GmbH).

Since subjects were speaking during the study, much care was taken to attenuate contamination of the EEG signal from facial muscles. EEG was first processed using the PREP pipeline, which performs bad channel identification/interpolation, re-referencing, and line-noise removal in a standardized, robust manner (Appelhoff et al., 2022; Bigdely-Shamlo et al., 2015). Subsequently, we bandpass-filtered the EEG between 1-70 Hz and then decomposed the data into 60 independent components (ICs) and removed components that were correlated with the EOG channels. We also, importantly, removed ICs that showed evidence of facial EMG contamination using the spectral slope, peripheral (i.e. far from vertex) power, and focal point (i.e. low spatial smoothness) criteria previously validated by Dharmaprani and colleagues using data from paralyzed patients (Dharmaprani et al., 2016). After removal of artifactual ICs, we then epoched the data and dropped trials that still showed evidence of strong non-neural signal contamination (i.e. peak-to-peak voltage exceeding 150 microvolts). To address any subtler movement artifacts, we then computed the current source density transformation of the EEG data using a spherical spline surface Laplacian. This transformation, effectively a spatial high pass filter, attenuates the effects of volume conduction across the scalp such that each electrode predominantly reflect nearby physiological sources (Kayser & Tenke, 2015). Previous work, again using data from paralyzed patients, has shown that the surface Laplacian effectively removes EMG contamination from central and peri-central electrodes (Fitzgibbon et al., 2013). Finally, the EEG was down-sampled to 140 Hz (i.e. twice the lowpass cutoff) to reduce the computational burden of our analysis.

In addition to the preprocessed data itself, detailed preprocessing reports for each individual subject are available as part of our open dataset. These reports contain power spectra of the EEG signal after PREP, a list of electrodes that were interpolated (and the criteria by which they were flagged during PREP), visualized scalp topographies of all removed ICs, and remaining trial counts for each condition after rejection criteria are applied.

5. EGG and Audio Acquisition and Synchronization

EGG was acquired using an EG2-PCX2 electroglottograph amplifier (Glottal Enterprises, Inc.) and audio using a Behringer ECM-8000 microphone with a roughly flat frequency response. These signals were sent from the EGG amplifier to a StimTrak (Brain Products, GmbH) via an audio passthrough cable, and were then digitized by our EEG amplifier so that all signals were recorded on the same clock. Like most EGG systems, the EG2-PCX2 has a built-in highpass filter, which introduces phase-specific delays into the recorded signal. To reverse this phase delay, we applied a digital equivalent of this 20 Hz highpass filter backwards after acquisition. (This sequential forward-backward filtering is the same operation that "zero phase shift" filters, like *scipy*'s and MATLAB's *filtfilt* functions, apply.)

6. Data and Code Availability

All raw data and preprocessed derivatives, included fitted models, are available on *OpenNeuro* (https://doi.org/10.18112/openneuro.ds005403.v1.0.1) and are organized according to the Brain Imaging Data Structure conventions for EEG (Pernet et al., 2019). Versioned analysis code is archived on *Zenodo* (https://zenodo.org/doi/10.5281/zenodo.13238912), and task code is available on GitHub (https://github.com/apex-lab/daf-experiment/releases/tag/v1.0.0).

References

Alonso, R. G., Trevisan, M. A., Amador, A., Goller, F., & Mindlin, G. B. (2015). A circular model for song motor control in Serinus canaria. *Frontiers in Computational Neuroscience*, *9*. https://doi.org/10.3389/fncom.2015.00041

Amador, A., Perl, Y. S., Mindlin, G. B., & Margoliash, D. (2013). Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature*, *495*(7439), Article 7439. https://doi.org/10.1038/nature11967

Appelhoff, S., Hurst, A., Lawrence, A., Li, A., Mantilla, R., Yorguin, J., O'Reilly, C., Xiang, L., & Dancker, J. (2022). *PyPREP: A Python implementation of the preprocessing pipeline (PREP) for EEG data* [Computer software]. https://doi.org/10.5281/zenodo.6363575

Ashmore, R. C., Renk, J. A., & Schmidt, M. F. (2008). Bottom-Up Activation of the Vocal Motor Forebrain by the Respiratory Brainstem. *Journal of Neuroscience*, *28*(10), 2613–2623. https://doi.org/10.1523/JNEUROSCI.4547-07.2008

Ashmore, R. C., Wild, J. M., & Schmidt, M. F. (2005). Brainstem and Forebrain Contributions to the Generation of Learned Motor Behaviors for Song. *Journal of Neuroscience*, *25*(37), 8543–8554. https://doi.org/10.1523/JNEUROSCI.1668-05.2005

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://www.frontiersin.org/articles/10.3389/fninf.2015.00016

Boari, S., Perl, Y. S., Amador, A., Margoliash, D., & Mindlin, G. B. (2015). Automatic reconstruction of physiological gestures used in a model of birdsong production. *Journal of Neurophysiology*, *114*(5), 2912–2922. https://doi.org/10.1152/jn.00385.2015

Bradshaw, A. R., Lametti, D. R., & McGettigan, C. (2021). The Role of Sensory Feedback in Developmental Stuttering: A Review. *Neurobiology of Language*, *2*(2), 308–334. https://doi.org/10.1162/nol_a_00036

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., Resnik, P., & Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, *12*, e85012. https://doi.org/10.7554/eLife.85012

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*. https://doi.org/10.3389/fnhum.2016.00604

Daou, A., & Margoliash, D. (2020). Intrinsic neuronal properties represent song and error in zebra finch vocal learning. *Nature Communications*, *11*(1), 952. https://doi.org/10.1038/s41467-020-14738-7

Dharmaprani, D., Nguyen, H. K., Lewis, T. W., DeLosAngeles, D., Willoughby, J. O., & Pope, K. J. (2016). A comparison of independent component analysis algorithms and measures to discriminate between EEG and artifact components. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 825–828.

Dima, G. C., Copelli, M., & Mindlin, G. B. (2018). Anticipated Synchronization and Zero-Lag Phases in Population Neural Models. *International Journal of Bifurcation and Chaos*, *28*(08), 1830025. https://doi.org/10.1142/S0218127418300252

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. https://doi.org/10.1016/j.neuroimage.2022.119728

Fetterman, G. C., & Margoliash, D. (2023). *Rhythmically bursting songbird vocomotor neurons are organized into multiple sequences, suggesting a network/intrinsic properties model encoding song and error, not time* (p. 2023.01.23.525213). bioRxiv. https://doi.org/10.1101/2023.01.23.525213

Fitzgibbon, S. P., Lewis, T. W., Powers, D. M. W., Whitham, E. W., Willoughby, J. O., & Pope, K. J. (2013). Surface Laplacian of Central Scalp Electrical Signals is Insensitive to Muscle Contamination. *IEEE Transactions on Biomedical Engineering*, *60*(1), 4–9. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2012.2195662

Gibb, L., Gentner, T. Q., & Abarbanel, H. D. I. (2009). Brain Stem Feedback in a Computational Model of Birdsong Sequencing. *Journal of Neurophysiology*, *102*(3), 1763–1778. https://doi.org/10.1152/jn.91154.2008

Golfinopoulos, E., Tourville, J. A., & Guenther, F. H. (2010). The integration of large-scale neural network modeling and functional brain imaging in speech motor control. *NeuroImage*, *52*(3), 862–874. https://doi.org/10.1016/j.neuroimage.2009.10.023

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Harutyunyan, A., Dabney, W., Mesnard, T., Gheshlaghi Azar, M., Piot, B., Heess, N., van Hasselt, H. P., Wayne, G., Singh, S., Precup, D., & Munos, R. (2019). Hindsight Credit Assignment. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html

Howell, P., & Archer, A. (1984). Susceptibility to the effects of delayed auditory feedback. *Perception & Psychophysics*, *36*(3), 296–302. https://doi.org/10.3758/BF03206371

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*(7600), 453–458. https://doi.org/10.1038/nature17637

IEEE Recommended Practice for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. IEEE Transactions on Audio and Electroacoustics. https://doi.org/10.1109/TAU.1969.1162058

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., & Lizarazu, M. (2024). On the speech envelope in the cortical tracking of speech. *NeuroImage*, *297*, 120675. https://doi.org/10.1016/j.neuroimage.2024.120675

Joris, P. X., Smith, P. H., & Yin, T. C. T. (1998). Coincidence Detection in the Auditory System: 50 Years after Jeffress. *Neuron*, *21*(6), 1235–1238. https://doi.org/10.1016/S0896-6273(00)80643-1

Kakade, S. M. (2001). A Natural Policy Gradient. *Advances in Neural Information Processing Systems*, *14*. https://proceedings.neurips.cc/paper_files/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html

Kayser, J., & Tenke, C. E. (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology. *International Journal of Psychophysiology*, *97*(3), 171–173. https://doi.org/10.1016/j.ijpsycho.2015.06.001

Kim, K. S., Wang, H., & Max, L. (2020). It's About Time: Minimizing Hardware and Software Latencies in Speech Research With Real-Time Auditory Feedback. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2522–2534. https://doi.org/10.1044/2020_JSLHR-19-00419

Legaspi, R., & Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-12170-0

Matell, M. S., & Meck, W. H. (2004). Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cognitive Brain Research*, *21*(2), 139–170. https://doi.org/10.1016/j.cogbrainres.2004.06.012

Moll, F. W., Kranz, D., Corredera Asensio, A., Elmaleh, M., Ackert-Smith, L. A., & Long, M. A. (2023). Thalamus drives vocal onsets in the zebra finch courtship song. *Nature*, *616*(7955), 132–136. https://doi.org/10.1038/s41586-023-05818-x

Mulliken, G. H., Musallam, S., & Andersen, R. A. (2008). Forward estimation of movement state in posterior parietal cortex. *Proceedings of the National Academy of Sciences*, *105*(24), 8170–8177. https://doi.org/10.1073/pnas.0802602105

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, *5*(11), eaay6279. https://doi.org/10.1126/sciadv.aay6279

Oldfield, R. C. (2013). *Edinburgh Handedness Inventory* [Dataset]. https://doi.org/10.1037/t23111-000

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Perl, Y. S., Arneodo, E. M., Amador, A., Goller, F., & Mindlin, G. B. (2011). Reconstruction of physiological instructions from Zebra finch song. *Physical Review E*, *84*(5), 051909. https://doi.org/10.1103/PhysRevE.84.051909

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, *6*(1), 103. https://doi.org/10.1038/s41597-019-0104-8

Schneider, D. M., & Mooney, R. (2018). How Movement Modulates Hearing. *Annual Review of Neuroscience*, *41*(Volume 41, 2018), 553–572. https://doi.org/10.1146/annurev-neuro-072116-031215

Schneider, D. M., Sundararajan, J., & Mooney, R. (2018). A cortical filter that learns to suppress the acoustic consequences of movement. *Nature*, *561*(7723), 391–395. https://doi.org/10.1038/s41586-018-0520-5

Sharpee, T. O., Nagel, K. I., & Doupe, A. J. (2011). Two-dimensional adaptation in the auditory forebrain. *Journal of Neurophysiology*, *106*(4), 1841–1861. https://doi.org/10.1152/jn.00905.2010

Simonyan, K., & Horwitz, B. (2011). Laryngeal Motor Cortex and Control of Speech in Humans. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *17*(2), 197–208. https://doi.org/10.1177/1073858410386727

Sitt, J. D., Amador, A., Goller, F., & Mindlin, G. B. (2008). Dynamical origin of spectrally rich vocalizations in birdsong. *Physical Review E*, *78*(1), 011905. https://doi.org/10.1103/PhysRevE.78.011905

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their

responses to natural stimuli. *Network: Computation in Neural Systems*, *12*(3), 289–316. https://doi.org/10.1080/net.12.3.289.316

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America*, *83*(4), 1536–1552.

Titze, I. R. (1998). Measurements for voice production: Research and clinical applications. *J. Acoust. Soc. Am*, *104*, 1148.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*(12), Article 12. https://doi.org/10.1038/nrn3112

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11), Article 11. https://doi.org/10.1038/81497

Yamamoto, K., & Kawabata, H. (2014). Adaptation to delayed auditory feedback induces the temporal recalibration effect in both speech perception and production. *Experimental Brain Research*, *232*(12), 3707–3718. https://doi.org/10.1007/s00221-014-4055-1