Modeling sensorimotor adaptation in speech through alterations to forward and inverse models

Taijing Chen[1], *Adam Lammert*[2], *Benjamin Parrell*[3]

[1]Department of Computer Sciences, University of Wisconsin-Madison, USA
[2]Department of Biomedical Engineering, Worcester Polytechnic Institute, USA
[3]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, USA

tchen284@wisc.edu, alammert@wpi.edu, bparrell@wisc.edu

Abstract

When speakers are exposed to auditory feedback perturbations of a particular vowel, they not only adapt their productions of that vowel but also transfer this change to other, untrained, vowels. However, current models of speech sensorimotor adaptation, which rely on changes in the feedforward control of specific speech units, are unable to account for this type of generalization. Here, we developed a neural-network based model to simulate speech sensorimotor adaptation, and assess whether updates to internal control models can account for observed patterns of generalization. Based on a dataset generated from the Maeda plant, we trained two independent neural networks: 1) an inverse model, which generates motor commands for desired acoustic outcomes and 2) a forward model, which maps motor commands to acoustic outcomes (prediction). When vowel formant perturbations were given, both forward and inverse models were updated when there was a mismatch between predicted and perceived output. Our results replicate behavioral experiments: the model altered its production to counteract the perturbation, and showed gradient transfer of this learning dependent on acoustic distance between training and test vowels. These results suggest that updating paired forward and inverse models provides a plausible account for sensorimotor adaptation in speech.

Index Terms: sensorimotor adaptation, modeling, transfer of learning, internal models

1. Introduction

When speakers are exposed to alterations of their auditory feedback that perturb their vowel formants, they respond by adapting their speech over time to oppose that perturbation. For example, a perturbation which raises the first vowel formant (F1) causes participants to produce lower F1 values [1, 2]. Previous work has shown that such speech motor learning transfers in a gradient fashion to untrained words and vowels [3]. In this study, participants produced a word (/pVn/) containing one of the vowels /ɪ/, /ɛ/, or /æ/ repeatedly over many trials. While speaking, participants received altered auditory feedback, such that the F1 they heard through headphones was higher than the F1 they produced. As expected, participants in all cases adapted by lowering their F1 by the end of training. Subsequently, all participants produced the word /pɛn/ with unaltered feedback in order to assess the transfer of learning from the trained vowel (/ɪ/, /ɛ/, or /æ/) to /ɛ/. The results showed that learning did transfer, but only partially. The magnitude of transfer was related to the acoustic distance (in F1-F2 space) between the training and transfer vowels; transfer was higher for /æ/ than for /ɪ/, and both were lower than the amount of learning retained in /ɛ/.

While this transfer of learning has been demonstrated experimentally, current models of speech sensorimotor adaptation cannot account for such generalization. To date, the most well-developed model of adaptation is found in the DIVA (Directions Into Velocities of Articulators) model [4, 5]. In the DIVA model, speech motor control starts with the selection of atomic "speech sounds," planning units with 1) motor trajectories and 2) trajectories of expected auditory and somatosensory feedback. Auditory errors (mismatches between expected and perceived auditory feedback) result in feedback motor commands that correct speech in real time. These feedback commands are subsequently used to update the stored motor trajectory for that speech sound, driving adaptation. However, since speech units are independent of each other in DIVA, any learned adaptation to one speech sound is not predicted to transfer to other units. To induce meaningful generalization, the model would need to either infer some relationships among particular speech units or learn more general mappings between motor goals, motor behavior, and sensory output.

The present paper aims to establish the validity of an alternative model of sensorimotor adaptation that can computationally account for the generalization pattern observed in laboratory settings. It has been suggested that sensorimotor adaptation may arise from changes to internal models; either forward models that predict the sensory consequences of actions (e.g., [6, 7]), inverse models (or, more generally, control policies) that select appropriate motor commands for desired movements [8], or both [9]. Motivated by this paradigm, we developed a simple model of speech production that combines forward and inverse models. The inverse model generates motor commands based on desired acoustic targets, while the forward model generates a prediction of the acoustic outcomes of those motor commands, which is used to determine the occurrence of auditory errors.

To investigate the feasibility of this approach, we assessed the ability of this model to reproduce basic behavioral results of sensorimotor adaptation and its pattern of transfer. We implemented the forward model and the inverse model as two independent neural networks given the well-established ability of these networks to learn arbitrary input-output relationships. We proceeded to conduct two simulation studies to test the response of the model to an applied perturbation of the perceived vowel formant frequencies, and observed results consistent with established behavioral outcomes.

2. Methods

2.1. Model architecture

Our model has three components (see Fig 1): a plant based on the Maeda vocal tract model [10], an inverse model which gen-

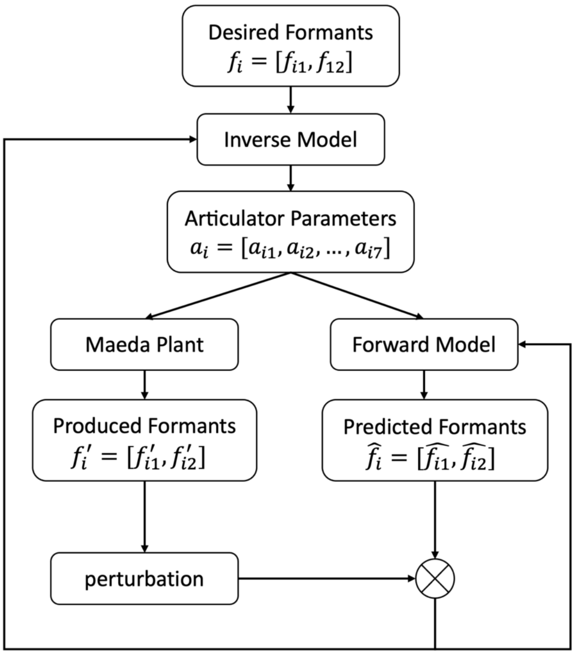

Figure 1: *Model Architecture.*

erates Maeda model parameters that correspond to desired F1 and F2 frequencies, and a forward model which predicts the F1 and F2 frequencies resulting from a particular set of Maeda parameters. The plant takes in 7 articulatory parameters that define the 2D shape of the vocal tract in the midsagittal plane and outputs 5 corresponding formants. Articulatory parameters are based on the primary principle components of observed vocal tract deformations in an x-ray dataset of speech [10].

On each trial, the model takes in an acoustic goal (the vector of formant frequencies f_i), representing the desired F1 and F2 values associated with vowel i. Given this acoustic goal, the inverse model generates motor commands (a vector of articulatory parameters a_i) capable of completing this task, which is passed to both the plant (i.e., the "motor command") and the forward model (i.e., "efference copy"), with random additive noise $n_i \sim N(0, 0.05)$ to simulate motor and prediction noise, respectively. From this input, the plant produces the formants f_i' corresponding to the input articulator values (i.e., speech production). At the same time, the forward model predicts the formant output of the plant ($\hat{f}_i$).

2.2. Model training dataset

In order to train the initial forward and inverse models described above, we first generated a dataset of Maeda parameter values and the corresponding formant frequencies. Using the plant, we generated pairs of motor commands (vectors of articulatory parameters, $a_i = [a_{i1}, a_{i2}, ...a_{i7}]$) and their and their corresponding acoustic outcomes (first two formant values, ($f_i = [f_{i1}, f_{i2}]$). Each of the 7 Maeda control parameters was uniformly sampled between -3.0 and 3.0, which corresponds to ±3 standard deviations of each principle component of movement relative to the x-ray data used to build the plant model [10]. This dataset was restricted to only those combinations of parameter values with vowel-like acoustics, such that a synthesized pair of articulatory parameters and vowel formants was only included in the dataset if its articulatory parameters resulted in a complete output of all of the first five formant values. To remove outliers, pairs whose formant values fell further than 3 standard deviations from the F1 or F2 mean were excluded. The dataset used for training is shown in Figure 2.

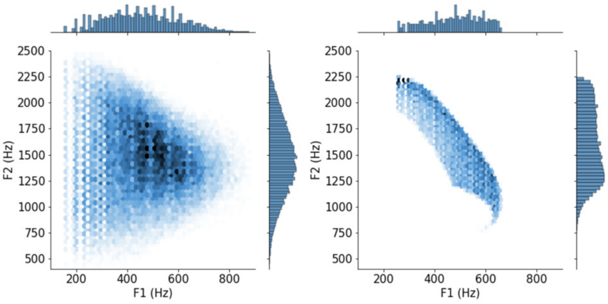

Figure 2: *Dataset. Forward Model (left); Inverse Model, where only the "jaw" and "tongue" parameters were free to vary (right).*

Table 1: *Forward Model Accuracy. F1 range: 152-878 Hz; F2 range 400-2496 Hz.*

Accuracy	**F1 (Hz)**	**F2 (Hz)**
mean	5.251	10.620
std	6.880	16.373
25%	1.543	2.956
50%	3.301	6.468
75%	6.144	12.159

2.3. Forward and Inverse models

Before we could examine adaptation, the forward and inverse models first need to learn the relationship between the articulatory parameter values and the resulting formants, analogous to human participants' prior knowledge before the adaptation experiments. In our architecture, the forward model and the inverse model were trained as two independent multi-layer neural networks. The forward training dataset described above (about 58,000 pairs of data) was used to tune the forward model to learn to predict the formant values associated with particular Maeda parameter sets. The average absolute prediction error of the forward model is about 5.25 Hz for F1 and about 10.62 Hz for F2 (Table 1).

The goal of the inverse model is to generate articulator parameters that result in desired formant values. However, the mapping from vowel formants to articulatory parameters in this plant model is ill-posed (i.e., the often-cited "many-to-one" mapping problem [11, 12]), making it difficult to build an accurate mapping between the complete input (formants $F \in \mathbb{R}^2$) and output (articulators $A \in \mathbb{R}^7$) space. In order to make this mapping more tractable for the current proof-of-concept work, we constructed an inverse model that maps the formant values to articulator parameters in only a restricted subspace of the possible outputs with equivalent dimensionality to the input space. To that end, we selected a pair of articulatory parameters most plausibly related to tongue shape control for vowel production. Previous studies on speech motor adaptation and transfer tested human participants mainly on productions of front vowels [13, 3]. Plotting the formants generated by all potential pairs

Table 2: *Inverse Model Accuracy. Range for both jaw and tongue parameters is -3.0 to 3.0.*

Accuracy	jaw	tongue
mean	0.124	0.092
std	0.093	0.069
25%	0.050	0.040
50%	0.107	0.079
75%	0.178	0.128

of articulatory parameters with relatively equivalent error rates, we found that the combination of the *jaw* and *tongue* parameters produced formants covering the front vowel region and were thus most suitable for the current modeling task. Thus, in our final implementation, the inverse model was trained on pairs of 1) formant frequency vectors and 2) articulatory vectors where the *jaw* and *tongue* parameters were sampled randomly, with all other parameters set to 0.

In addition, although the articulator parameters in the training data ranged from -3.0 to 3.0, we tuned an L2-regularizer to penalize large articulator outputs during training (i.e., including squared magnitudes of the outputs in the cost function) [14]. Conceptually, this regularization essentially produces the smallest movements that could result in a particular vowel formant pattern, consistent with the idea that the speech motor system minimizes articulatory effort [4, 5]. The absolute prediction errors of the inverse model are about 0.12 for the *jaw* articulator parameter and about 0.09 for the *tongue* articulator parameter.

2.4. Simulating adaptation and transfer

Here, we assess the ability of our model to simulate: 1) adaptation to an auditory perturbation that alters vowel formants and 2) transfer of that adaptation to untrained vowels.

To simulate adaptation, we perturbed the F1 output of the plant. On each iteration, after the plant produced true formants f_i', we added a perturbation vector $p = [75, 0]$ to this output. This "perceived" feedback, $f_i' + p$, and Maeda motor commands a_i were used to update both the forward and the inverse models. We note here that the magnitude of the perturbation we used is somewhat smaller than the one used by [3] (25% of vowel F1). This was necessary as the restricted model we used, with only two free articulator parameters, covers only a portion of the full vowel space (Figure 2). Crucially, we wanted to ensure that the perturbed formant values $f_i' + p$ would fall within the region covered by the training dataset to ensure the inverse model's accuracy. If the error detection gate detected an error between predicted formants $\hat{f}_i$ and the produced formants f_i' (threshold $= 5.88$Hz, i.e., mean F1 prediction error of the forward model), the forward model was updated by the pair $(a_i, f_i' + p)$ and the inverse model would learn from the equivalent pair $(f_i' + p, a_i)$. This updated inverse model was then used on the following iteration to generate the next articulator vector a_{i+1} for the (invariant) acoustic goal f_{i+1}. Both forward and inverse models were re-trained during these adaptation iterations with learning rates 1/10 of the original rates. We additionally added a restriction to this adaptation process to constrain the inverse model's output articulator parameters. Recall that the articulatory parameters in the plant normally range from -3.0 to 3.0. If, after an update, the output articulator parameters of the inverse model a_{i+1} fell out of this range, the model would discard this a_{i+1}, roll back to the previous state, and generate a new articulatory vector $a_{i+1}(new)$ instead. Because the articulatory vector a_i had random noise n_i on each iteration, even if $inverse\_model_i$ was the same as $inverse\_model_{i+1}$, the pair $(a_i, f_i' + p)$ would be different from $(a_{i+1}, f_{i+1}' + p)$. In other words, the architecture would not be repeatedly trained on the same data pair. These roll-backs occurred on a very small proportion of iterations ($< 0.1\%$).

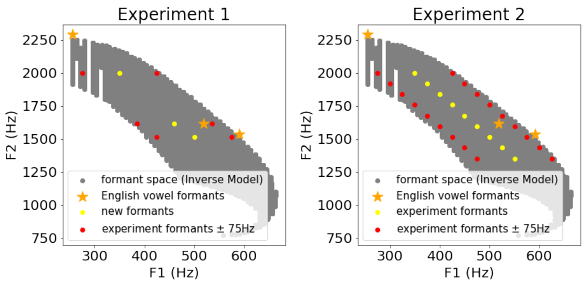

Figure 3: *Experiment Formants.*

We ran two simulation studies to test the architecture's response to perturbation and the effects on transfer. In *Experiment 1*, we aimed to replicate the behavioral results of [3]. To reiterate, this experiment perturbed participants' auditory feedback on their productions of vowels /ɪ/, /ɛ/, and /æ/ and tested transfer of learning on vowel /ɛ/. Since the formant frequencies corresponding English vowels /ɪ/, /ɛ/, and /æ/ in the plant are on the edge of the F1-F2 vowel space covered by the inverse model's training dataset, to ensure $f_i' \pm p$ is within this region, we heuristically defined a line l that is furthest from the region's left and right edges. On this line, we picked three (f_1, f_2) points closest to the F1-F2 positions of real English vowel /ɪ/, /ɛ/, and /æ/ as the substitute acoustic goals (Figure 3). We use IH', EH' and AE', respectively, to denote these targets. In each condition, the model produced one of IH', EH' and AE' (training vowels) as the acoustic goal, with the +75 Hz perturbation applied to F1. This adaptation process was iterated until the forward and inverse models showed convergence on a stable mapping (change in loss less than 1e-4). After the internal models fully adapted to the applied perturbation, we tested the overall model's 1) final "behavioral" adaptation to the training vowels and 2) the transfer of learning to EH'. For each training condition (IH', EH', AE'), we ran 30 adaptation simulations. To ensure the results from *Experiment 1* are not caused by the specific substitute vowels we used, we ran *Experiment 2*, where we discretized the line l into nine F1-F2 training points $(v_1, v_2, ..., v_9)$ and tested the effects of transfer on the middle vowel (v_5).

In each experiment, we repeated each condition (training on IH', EH', or AE') 30 times. The magnitude of adaptation was measured as the difference between the acoustic targets (which remained invariant throughout each simulation) and the unperturbed formant output of the Maeda plant after training. Transfer of learning was measured as change in formant outputs for EH' after training, expressed as a proportion of the final adaptation of the training targets (IH', EH', or AE'). The acoustic similarity between vowels was measured by their Euclidean distance in F1-F2 space, following [3].

3. Results

We observed gradient transfer of learning in both experiments (Figure 4 & 5). In all conditions, average transfer of learning

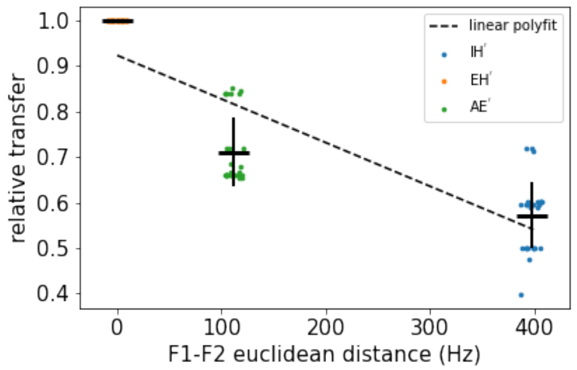

Figure 4: *Experiment 1 Results.*

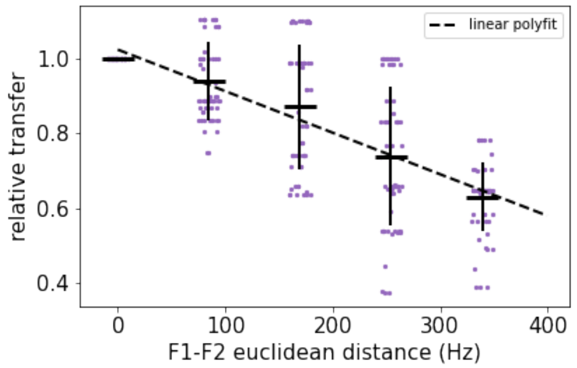

Figure 5: *Experiment 2 Results.*

was smaller when the training vowel is farther away from the testing vowel. In Experiment 1, less transfer was seen for IH' (0.57 ± 0.07) than for AE' (0.71 ± 0.08); both were less than EH' (1.00 ± 0.00). All differences were highly significant (overall main effect of distance: $F(2, 87) = 381.8, p < 0.0001$; all comparisons $p < 0.0001$). We additionally calculated the correlation between the magnitude of transfer and the acoustic distance between trainer and test vowels. When the correlation was computed from the data of each simulation, transfer of learning is reliably correlated with the Euclidean distances between the training and testing vowels on the F1-F2 space ($r = 0.853, p < 0.01$). We did not conduct a similar correlation using the computed from means of the three vowels, as this would result in a correlation based on only 3 datapoints. In Experiment 2, we saw a similar correlation between the transfer of learning and training-testing vowel distances ($r = 0.718, p < 0.01$ when he correlation was computed from the data of each trial; $r = 0.988, p < 0.01$ when the correlation was computed from the vowel mean).

4. Discussion

Our current results are based on a highly simplified model and serves primarily as a proof of concept for this approach. Further work is needed to assess the ability of this approach to account for sensorimotor adaptation in a more realistic model. First, one main limitation of the current architecture is the constraints imposed on the inverse model. In order to achieve reasonable accuracy in the inverse model, we reduced the controllable parameters of the Maeda model to only the *jaw* and *tongue* parameters, resulting in a reduced vowel space. Better ways of calculating the inverse model are required expand this to the full set of Maeda parameters and the complete vowel space. One potential alternative is the invertible neural network (INN) [15]. Instead of solving the forward process and the inverse process independently, INNs use additional latent variables to capture information lost in the forward process. Owing to its invertible architecture, by training the forward process, the inverse process can be retrieved. However, modeling speech motor learning using INN would assume the forward and the inverse processes are closely related, which which may not be the case [8].

Second, our current approach uses a fairly simple process to update internal models. In the current architecture, the output of the forward model is compared with the (potentially perturbed) output of the plant. When a discrepancy is detected above the intrinsic error in the forward model, the paired motor command and plant acoustics are gated to update both forward and inverse models. However, this process is likely substantially more complex in the real world. Learning is likely dependent on other statistical information (such as uncertainty about the source of sensory errors, e.g., [16]); this could be implemented by the using the forward model's prediction to tune the learning rate and step size of model adaptation. The architecture could also be expanded to integrate additional components and better reflect the experience of human speakers. For example, real speakers receive not only auditory but also somatosensory feedback. Notably, conflicts between these sensory modalities have been suggested to underlie the partial adaptation typical in human speech adaptation [17], a feature absent from our current simulations which fully adapted to the applied perturbation.

Third, both forward and inverse models are updated during the learning phase in our current model. However, the exact role each of these models plays in human sensorimotor adaptation remains unclear. It is possible that in human speech motor learning, only one of the models is updated or that they are updated at different rates. If only the forward model is updated, we hypothesize that no behavioral changes will be observed (since the selection of motor commands by the inverse model would remains unchanged); if we only updated the inverse model, we hypothesize that the model would be less stable, as learning would no longer be appropriately gated by accurate sensory predictions. Such predictions remain to be tested.

Lastly, our current model generates a single time point for each trial (essentially, assuming that vowel production is time-invariant). Of course, real speech varies substantially over time. Our current approach could be easily incorporated into models of online speech production that rely on state feedback control [18], such as the FACTS model [19]. These models similarly rely on both forward models and inverse models (or, more generally, control policies) for motor control; these internal models could be updated in a similar manner as that proposed here.

Despite these limitations, the current results suggest sensorimotor adaptation in speech could plausibly arise from updates to internal forward and inverse models. Updating such models leads to adaptation of the trained vowel, as expected. These changes in behavior also transfer to untrained vowels as a function of the acoustic distance between training and test vowels, consistent with behavioral results in human speakers.

5. References

[1] J. F. Houde and M. I. Jordan, “Sensorimotor adaptation in speech production,” *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[2] D. W. Purcell and K. G. Munhallr, “Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation,” *Journal of Acoustical Society of America*, vol. 120, no. 2, pp. 966–977, 2006.

[3] A. Rochet-Capellan, L. Richer, and D. J. Ostry, “Nonhomogeneous transfer reveals specificity in speech motor learning,” *Journal of Neuroscience*, vol. 107, no. 6, pp. 1711–1717, 2012.

[4] J. A. Tourville and F. H. Guenther, “The diva model: A neural theory of speech acquisition and production,” *Language and Cognitive Processes*, vol. 26, pp. 952–981, 2011.

[5] F. H. Guenther, *Neural control of speech*. The MIT Press, 2016.

[6] A. J. Bastian, “Learning to predict the future: The cerebellum adapts feedforward movement control,” *Curr Opin Neurobiol*, vol. 16, no. 6, p. 645–649, 2006.

[7] J. R. Flanagan, P. Vetter, R. S. Johansson, and D. M. Wolpert, “Prediction precedes control in motor learning,” *Current Biology*, vol. 13, no. 2, p. 146–150, 2003.

[8] A. M. Hadjiosif, J. W. Krakauer, and A. M. Haith, “Did we get sensorimotor adaptation wrong? implicit adaptation as direct policy updating rather than forward-model-based learning,” *Journal of Neuroscience*, 2021.

[9] D. Wolpert and M. Kawato, “Multiple paired forward and inverse models for motor control,” *Neural Networks*, vol. 11, no. 7, pp. 1317–1329, 1998.

[10] S. Maeda, “A digital simulation method of the vocal-tract system,” *Speech Communications*, vol. 1, no. 3, pp. 199–229, 1982.

[11] B. Atal, J. Chang, M. Mathews, and J. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *The Journal of the Acoustical Society of America*, vol. 63, no. 5, p. 1535—1553, May 1978.

[12] C. Qin and M. A. Carreira-Perpiñán, “An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping.” INTERSPEECH, 01 2007, pp. 74–77.

[13] A. Rochet-Capellan and D. J. Ostry, “Simultaneous acquisition of multiple auditory–motor transformations in speech,” *Journal of Neuroscience*, vol. 31, no. 7, pp. 2657–2662, 2011.

[14] V. N. Sorokin, A. S. Leonov, and A. V. Trushkin, “Estimation of stability and accuracy of inverse problem solution for the vocal tract,” *Speech Communication*, vol. 30, no. 1, pp. 55–74, 2000.

[15] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Köthe, “Analyzing inverse problems with invertible neural networks,” in *International Conference on Learning Representations*, 2019.

[16] K. Wei and K. Körding, “What drives motor adaptation?” *J Neurophysiol*, vol. 101, no. 2, p. 655–664, 2009.

[17] D. R. Lametti, S. M. Nasir, and D. J. Ostry, “Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback,” *Journal of Neuroscience*, vol. 32, no. 27, p. 9351–9358, 2012.

[18] J. F. Houde and S. S. Nagarajan, “Speech production as state feedback control,” *Frontiers in Human Neuroscience*, vol. 5, p. 82, 2011.

[19] B. Parrell, V. Ramanarayanan, S. Nagarajan, and J. Houde, “The facts model of speech motor control: Fusing state estimation and task-based control,” *PLOS Computational Biology*, vol. 15, no. 9, pp. 1–26, 09 2019.