Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Self-supervised Optimality-Guided Learning of Speech Articulation

Juraj Šimko[1], *Benjamin Elie*[2], *Alice Turk*[2]

[1]Department of Digital Humanities, University of Helsinki, Finland
[2]LEL, The University of Edinburgh, Scotland

juraj.simko@helsinki.fi, belie@ed.ac.uk, a.turk@ed.ac.uk

Abstract

This paper introduces a novel approach for modeling speech articulatory planning based on Optimal Control Theory. The presented approach uses an internal feed-forward controller model that learns to predict optimal articulatory commands minimizing a context-dependent objective function. This objective function combines conflicting tasks of minimizing articulatory effort and maximizing the recognition probability of a target vowel based on acoustic characteristics. We present a self-supervised optimality-guided architecture for training the feedforward internal model that directly uses the objective function as a training loss. Simulations involving isolated vowels of American-English show that online training of the internal model enables feedforward estimation of near-optimal articulatory parameters.

Index Terms: Speech acquisition, Articulatory Planning, Optimal Control Theory, Self-supervised learning

1. Introduction

A growing number of studies have proposed to use the *Optimal Control Theory* (OCT) approach [1, 2] to explain both phonological invariance and phonetic variability of speech production and articulatory planning of spoken utterances [3–13]. Besides speech articulation, OCT-based approaches have been applied to different types of bodily movements, including jumping [14], running [15], reaching and pointing [16, 17], eye [18] or arm movements [19]. OCT is now regarded as one of the dominant models of motor control of non-speech movements [20, 21].

Optimal control models traditionally assume that movements are planned to minimize a multi-objective cost function that penalizes movements not satisfying often mutually conflicting task requirements. Formally, the optimal movement specification minimizes the multi-objective cost function

$$J(\mathbf{x}; \mathbf{z}, \boldsymbol{\alpha}) = \sum^{N} \alpha_i C_i(\mathbf{x}; \mathbf{z}), \tag{1}$$

where C_i's and $\boldsymbol{\alpha} = \{\alpha_i\}$ are the cost functions and the weights associated with the individual task requirements, $\mathbf{x}$ is a vector specifying the movement, and $\mathbf{z}$ is a set of additional parameters. The trade-off weights α_i give more or less importance to each individual task, and might depend on context and other global requirements. The parameters $\mathbf{z}$ specify additional inputs of the planner, such as initial conditions of the system or the type of movement to be produced. The optimal movement

$$\mathbf{x}_{\text{optimal}} = \underset{\mathbf{x}}{\operatorname{argmin}} \left[J(\mathbf{x}; \mathbf{z}, \boldsymbol{\alpha}) \right] \tag{2}$$

minimizing the objective cost function thus depends on the weights and other specifications of the intended movement.

For speech, following Lindblom's hypothesis [22] that the observed phonetic variability is due to a parameterized trade-off between the economy of effort and intelligibility, several recent OCT models of speech production have used a continuous intelligibility measure as one of the components of multi-objective cost [5–13]. Other task requirements incorporated into the cost function in these studies included minimizing articulatory effort [4–7,9–13,23], achieving phonological [10,11,13] or spatio-temporal goals [5, 12], and sometimes also minimal duration [5, 7, 13]. The approach has successfully reproduced several well-known characteristics of speech, such as reduction and shortening of unstressed vowels [7, 13], acoustic and articulatory effects of Lombard speech [24], and consonantal quantity contrasts [25], making OCT a good candidate for explaining various aspects of speech motor planning.

Deriving optimal solutions of an OCT task, *i.e.*, finding the minima of an objective cost function for a given set of weights $\boldsymbol{\alpha}$ and parameters $\mathbf{z}$, imposes significant computational demands, in particular when the cost components represent highly non-linear relationships, such as mapping between articulation and intelligibility. If finding the optimal articulatory patterns is part of the speech planning process, it is important to address the question of how the optimization process might be implemented in the human cognitive system. A standard approach to this question replaces direct, online minimization of the objective cost function by gradual learning to estimate optimal solutions, using techniques such as dynamic programming [26] or reinforcement learning [27, 28].

In the case of speech, these questions have not been investigated, as current OCT-based models of speech articulatory planning use online minimization [5,7,12,13,24,29,30]: a minimization algorithm is used for each utterance, which eventually results in a plan that consists of the actual optimal solution. In this paper, we investigate an alternative approach where the optimal articulatory commands (movement parameter values) are estimated by a feedforward neural network. This optimality-predictor network is trained in a self-supervised manner using a set of task parameters as an input and the objective cost function directly as a loss. The proposed self-supervised optimality-guided learning paradigm thus does not require any speech data (e.g., pre-computed optimal solutions in the training database), nor any explicit cost minimization process during speech planning. Instead, the training process guarantees that the predicted articulatory commands are gradually, as the training goes on, better estimates of the actual optimal solutions as prescribed by the objective cost function / training loss.

Section 2 defines the OCT-based model of articulatory planning, and Section 3 its neural implementation using the optimality-guided training paradigm. To test the feasibility of the approach, results of simulations are presented in Section 4:

10.21437/Interspeech.2025-866

they use a minimal feedforward model to predict the optimal solution of a single movement per articulator for producing isolated American English vowels with specific trade-off weights.

2. A model of articulatory planning

Following [13], we use the Maeda model [31] for representing the geometry of the vocal tract. The Maeda articulatory model uses a set of seven parameters to represent the positions and movements of articulators in terms of jaw opening, height and position of the tongue dorsum, position of the tongue tip, lip opening, lip protrusion, and larynx height. The parameters have been derived from PCA, and are expressed in terms of standard deviations around the mean value of the principal components: these range between -3 and +3 standard deviations. In this paper, for simplicity, we assume only one movement per articulator and that the articulators are the seven parameters of the Maeda model. For any input articulatory position $\mathbf{x} = \{x_1, x_2, \ldots, x_7\}$, the Maeda-based acoustic model $\mathcal{A}$ produces a set of corresponding acoustic parameters $\mathbf{a} = \mathcal{A}(\mathbf{x})$.

The task of the articulatory planning system of our model is to find an optimal target articulatory position $\mathbf{x}_T$ that maximizes recognition probability of a given phoneme p and minimizes the articulatory effort of reaching the target $\mathbf{x}_T$ from a given starting position $\mathbf{x}_0$. Formally, for any given phoneme p, starting position $\mathbf{x}_0$ and weights α_E and α_R, the planner is computing a minimum of the objective function

$$J(\mathbf{x}; \mathbf{x}_0, p, \boldsymbol{\alpha}) = \alpha_E E(\mathbf{x}; \mathbf{x}_0) + \alpha_R R(\mathbf{x}; p), \qquad (3)$$

where $E(\mathbf{x}; \mathbf{x}_0)$ is the cost of articulatory effort for moving from $\mathbf{x}_0$), and $R(\mathbf{x}, p)$ is the recognition cost for p (the cost of not being intelligible based on acoustic characteristics). Note that, given the task to minimize the cost J, one of the weights is redundant, and for the remainder of this paper we set $\alpha_R = 1$.

The optimal end-point of the articulatory movement, i.e., the *planned movement target*, thus will be

$$\mathbf{x}_T = \underset{\mathbf{x}}{\operatorname{argmin}} \left[J(\mathbf{x}; \mathbf{x}_0, p, \boldsymbol{\alpha}) \right]. \qquad (4)$$

Following [24, 29], the *articulatory effort cost* is related to the amplitude of movement: it is the squared ℓ^2 norm of the amplitude vector defined as the difference between the position at the movement offset $\mathbf{x}$ and the position at the onset $\mathbf{x}_0$, hence:

$$E(\mathbf{x}; \mathbf{x}_0) = ||\mathbf{x} - \mathbf{x}_0||^2. \qquad (5)$$

The *recognition cost* is the cost of not being intelligible. It is related to intelligibility based on acoustic characteristics, and is defined here as the speaker's prediction of the probability of recognition of the target phoneme by the listeners, given the predicted acoustic consequences of the planned movements.

For a given phoneme p, the recognition cost is defined as

$$R(\mathbf{x}; p) = 1 - \mathrm{P}(p|\mathbf{a}) = 1 - \mathrm{P}(p|\mathcal{A}(\mathbf{x})), \qquad (6)$$

where $\mathrm{P}(p|\mathbf{a})$ is the probability of recognition of the target phoneme p given the acoustic $\mathbf{a}$ predicted from its position $\mathbf{x}$ by the Maeda-based model $\mathcal{A}$.

3. Neural implementation of the articulation planner

The overall architecture of our model is presented in Figure 1.

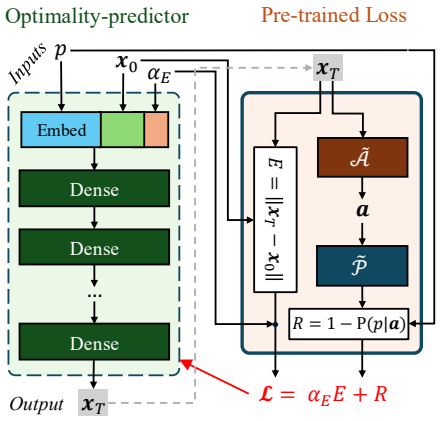

Figure 1: *Optimality-guided model architecture.*

3.1. Optimality-predictor

The core part of the model is the Optimality-predictor that is trained to predict (approximate) optimal end-points for an input consisting of the intended phoneme p, the weight α_E and the initial articulatory positions $\mathbf{x}_0$.

The Optimality predictor is implemented as a feed-forward neural network consisting of a stack of several dense layers. The categorical phoneme input is embedded, and the embedding is concatenated with the input value of α_E and with the input vector $\mathbf{x}_0$. The output layer size is adjusted to the required output size (i.e., seven in the simplified version presented here).

The predictor is trained using arbitrarily generated training and validation sets of input values in a self-supervised manner; that is, the training/validation database does not contain any target values of $\mathbf{x}_T$. Instead, we trained the predictor using a novel optimality-guided learning paradigm, where the training loss is computed to approximate the objective function in Eq. 3, using the predictor's output $\mathbf{x}_T$ and the values p, α_E and $\mathbf{x}_0$ passed on from the predictor's input.

3.2. Loss calculation

The loss for the predictor is calculated using two pre-trained networks approximating the two functions featuring in Eq. 6: (1) the Maeda-based articulation-to-acoustics mapping $\mathcal{A}$, and (2) the mapping of acoustics to phoneme probability P. The functions $\mathcal{A}$ and P in Section 2 model the *physics* behind transformation of articulatory positions to their acoustic consequences, and the *listener*'s categorical classification of acoustics in terms of phonemes, respectively. Their neural implementations, $\tilde{\mathcal{A}}$ and $\tilde{\mathrm{P}}$, on the other hand, play the roles of the internal approximations, used by the speaker for motor planning [1, 8, 32–34].

As shown in the lower part of Fig. 1, these two networks are chained to predict probability of recognition of the intended phoneme p, and estimate the value of the recognition cost R. This value is then summed with a value of the effort loss E multiplied by the weight α_E; the computations implemented using appropriate differentiable operators.

4. Simulations

In the implementation of this approach presented here, both neural models $\tilde{\mathcal{A}}$ and $\tilde{\mathrm{P}}$ are fully pre-trained before the start of the Optimality predictor training, and used purely to calculate the loss for training the predictor.

4.1. Training the internal articulatory-to-acoustic model and the internal acoustic perception model

In the implementation presented here, the acoustic vector $\mathbf{a}$ contains the predicted first four formant frequencies: $\tilde{\mathcal{A}}$ is a Multi-Layer Perceptron (MLP) that takes as its input a vector containing a static constellation of the 7 Maeda parameters and returns the predicted formant frequencies. The model is a regressor model with 5 layers of dimension (250, 140, 20, 40, 10), with *tanh* as the activation function of the hidden layers, and with a linear output layer which returns the predicted acoustic vector of dimension 4. The internal articulatory-to-acoustic model $\tilde{\mathcal{A}}$ was trained using randomly generated articulatory vectors $\mathbf{x}$ and the associated formant frequency vectors $\mathbf{a}$ precomputed using the Maeda model [31]. We generated 10^6 articulatory vectors following a uniform random distribution between -3 and +3 for the training set and 10^5 articulatory vectors following the same uniform distribution for the validation dataset. Note that the associated formant vector for a closed vocal tract was set to the null vector (all formants at 0). The model was trained using the AdaGrad optimizer with a learning rate of 10^{-3}, a batch size of 32, and training was stopped when the coefficient of determination (R^2 score) on the validation dataset was not improved for 5 successive epochs. At the end of training, the best R^2 score on the validation dataset was 0.986.

For the perception model, we used the pre-trained classifier model *american_english_11_vowels_vtr* from the PlanArt Python library [35]. This model is an extension of the 5-vowels perception model used in [13] to include the original 11 vowels of American English. The model was trained on real formant of American-English vowels derived from the Vocal Tract Resonance (VTR) corpus [36], using a Quadratic Discriminant Analysis (QDA) classifier to recognize 11 vowels of American-English based on the values of the first 4 formant frequencies.

4.2. Training the optimality-predictor

In this paper, we used four different optimality-predictor models (named M1, M2, M3, and M4) with the same architecture as shown in Fig 1 and whose characteristics are listed in Tab. 1. In order to investigate the effect of the dimension of the embedding layer used to encode the target phoneme, we trained 4 versions of M1, which differ from the dimension of the embedding layer, namely with embedding layer of size 4, 8, 16, and 32. The other models, M2, M3, and M4 have all an embedding layer of dimension 32. The activation function is *tanh* for all hidden layers and for all models. The output layer for all models is a linear layer that returns the estimated $\mathbf{x}_T$ vector containing the 7 target positions for each isolated vowel.

Table 1: *Dimension of hidden layers and number of trainable parameters of the different models used in this paper.*

Names	# layers	Dimension of layers	# parameters
M1	3	(40, 80, 40)	8799
M2	3	(64, 32, 64)	7623
M3	4	(64, 64, 64, 64)	15911
M4	5	(128, 64, 32, 64, 128)	27271

Training the optimality-guided articulatory planner consisted of simulating a learning phase: from a random initialization of the model, we used online training to train the model to produce sequences of vocalizations intended to be heard as sequences of isolated vowels. We chose online learning to simulate the speech acquisition phase of real speakers. The synaptic weights of the model were updated after each vowel production with a single update (corresponding to a training epoch with a single input). We used the AdaGrad optimizer with a learning rate of 10^{-3} shown to perform well for online training [37]. We trained our models using the *ssogl* submodule of the planart python library [35].

We randomly generated 10^6 input vectors, containing, for each of them, one effort weight α_E, one intended vowel phoneme p, and one initial-position vector of $\mathbf{x}_0$ containing random values of the 7 Maeda parameters. Random generation was performed using uniform distributions between 0 and 5 for α_E and between -3 and +3 for each Maeda parameter. The vowels were also taken at random from the set of 11 vowels using a discrete uniform distribution. The validation dataset was generated using the same process, but with only 100 samples. For the validation dataset we have also precomputed the optimal final-positions using the Nelder-Mead algorithm [23]. These optimal realizations were used purely to evaluate the performance of the optimality-predictor; no off-line optimization is part of the self-supervised training process.

5. Training results

5.1. Example of results for model M1

Fig. 2 shows the position of the vowels of the validation dataset returned by model M1 with an embedding layer of dimension 32, before training (just after initialization), and the same vowels after the last iteration of the learning process. At the early learning stage, the produced vowels are not distinguishable: they are all mixed around the center of the vowel space, each with large variance both in F1 and F2, as well as a large degree of overlap. At the late learning stage, the produced vowels are clearly separated, with no overlap (except for /ɚ/ and /ʊ/) with small variance in the F1 and the F2 dimensions. In addition, the vowels produced during the late learning stage are located in appropriate positions in the American-English vowel space, as observed in the VTR corpus [36]. The predicted cost of the validation dataset decreases as the number of produced vowels grows, which shows that the model learns to minimize the cost function through experience. At the end of the learning phase, the predicted cost of the validation dataset is close to the minimal cost (ground truth) estimated with full optimization (0.425 *vs.* 0.390), showing that the predictor has learned to almost predict the optimal solution of Eq. 4.

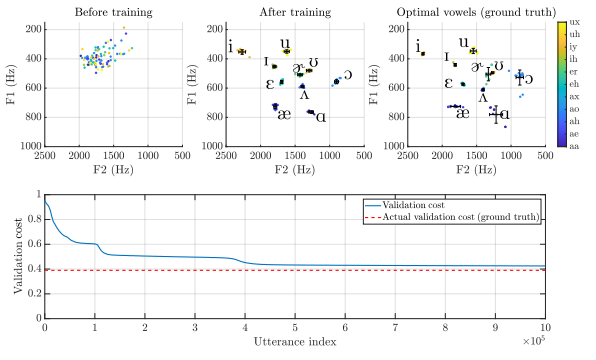

Figure 2: *The top panels show the positions of the predicted optimal vowels of each category for the validation dataset in F1-F2 space (left) before training, (center) after training consisting of 1 million productions, and (right) the actual optimal vowels. The bottom panel shows the evolution of the predicted cost of the validation dataset as a function of the number of vowel productions.*

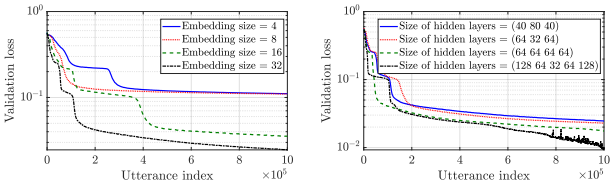

Figure 3: *Predicted cost of the validation dataset with respect to the baseline of (left) model M1 as a function of the embedding layer dimension, and (right) of the different architectures.*

5.2. Performance for different architectures

Fig. 3 shows the difference between the predicted cost (loss) of the validation dataset and the actual cost for different models. It shows that the dimension of the embedding layer impacts the predicted cost of the validation dataset, as it is reduced for larger numbers of embedding layers. Similarly, the complexity of the model impacts the predicted cost, as higher complexity also reduces the predicted cost.

5.3. Effect of the trade-off between low articulatory effort and high intelligibility

This section presents direct applications of speech production using our optimality-predictor models. For that purpose we create a test dataset that is made of 100 instances of each vowels with random initial positions. This dataset is repeated four times with different values of the weight α_E, namely $\alpha_E = 5$, $\alpha_E = 10$, $\alpha_E = 15$, and $\alpha_E = 20$. This is done to simulate different intended degrees of hyper- and hypo-articulated speech, going from a hyper-articulated speech scenario ($\alpha_E = 5$) to an intended hypo-articulated scenario ($\alpha_E = 20$) with values of α_E which have not been seen during training.

Fig. 4 represents the Vowel Space Density (VSD) when applying the 4 different models, with the same dimension of 32 for the embedding layer, on the test dataset. The VSD map represents the density (or number) of vowels for which formant values are sufficiently close to a given point in the map, as defined in [38]. In the hyper-articulated case (α_E=5, left column), the vowels are clearly separated as the VSD map shows clear modes of density. There are no clear differences across models. As the value of α_E increases, namely as the demand on low articulatory effort increases, vowels tend to merge together at a point located around the center of the vowel space. For $\alpha_E > 5$, Fig. 4 shows that the produced vowels shows discrepancies across used pre-trained models. For instance, for $\alpha_E = 10$, some models produce stronger vowel overlap (M1 and M3) than other models (M2 and M4). For $\alpha_E = 15$ and $\alpha_E = 20$, M2 produces vowels which are still slightly distinguishable as modes in the VSD map, while other models produce much stronger vowel overlap (especially M1 and M4).

6. Conclusions

This paper presents a minimal version of a new approach for Optimal Control based articulatory planning of spoken utterances. The approach presented here is based on the assumption that optimality is learned through experience. For that purpose, we model an articulatory planner as a fully connected feedforward Artificial Neural Network (ANN) that predicts optimal articulatory parameters given the context (e.g., intended speech style and intended phonemes): the synaptic weights of the ANN are updated after each production such that it learns to minimize the cost of the next production.

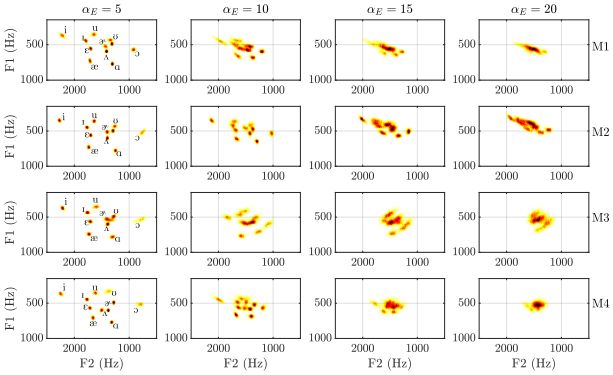

Figure 4: *Vowel density of the produced vowels by 4 different models (top to bottom) and for 4 different values of the weight α_E, namely 5, 10, 15, and 20, from left to right.*

Using online training, we successfully trained models to reduce the cost of production of vowels: the cost of production of a validation dataset (unseen during training) was found to be close to the actual cost after training. As a result, our simulations show that the optimality-guided approach is able to train a model to produce isolated vowels which acoustically correspond to actual vowels of American-English. In addition, the trained models predict the effect of different demands on low articulatory effort and high intelligibility: stronger importance on high intelligibility results in models producing hyper-articulated vowels while stronger importance of low articulatory effort results in models producing hypo-articulated vowels. The effect of the importance on low articulatory effort and high intelligibility on hypo- and hyper-articulation of vowels is qualitatively very similar to that reported in previous studies that consider full online optimization before each production [29, 30].

These promising results provide support for a new optimality-guided learning approach for speech acquisition. This offers new possibilities for investigating and identifying sources of speech variability. For instance, our simulations showed that different model architectures led to different patterns of vowel production. It could be also interesting to try different modifications of the model that reproduce known disorders of speech motor control.

For that purpose, the model architectures presented in this paper would need to be extended in several ways. First, we used pre-trained internal models to compute the cost function, assuming that the speaker had internal models that could already accurately predict the acoustic consequences of articulatory commands and that could predict the probability of recognition of phonemes given their predicted acoustic consequences. In practice, these models are likely to be learned online, as well. However, our approach is compatible with parallel online training of all internal models, including online training of sensory model after each production. This would require updates of the internal sensory models used to compute the cost function after each production. Since internal sensory models would stabilize after a sufficient number of produced utterances, the learning phase with stabilized internal sensory models would be equivalent to the learning phase presented in this paper.

Another major improvement would be to consider a sequence of phonemes. For that purpose, the architecture of the model could be modified to include recurrent connections in order to accurately process sequence of movements and phonemes to produce.

7. Acknowledgments

We gratefully acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A. Turk; Grant agreement No. 101019847).

8. References

[1] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[2] E. Todorov, "Optimal control theory," in *Bayesian brain: probabilistic approaches to neural coding*, D. K, Ed. MIT Press Cambridge, MA, 2006, pp. 268–298.

[3] W. L. Nelson, "Physical principles for economies of skilled movements," *Biological Cybernetics*, vol. 46, no. 2, pp. 135–147, 1983.

[4] W. L. Nelson, J. S. Perkell, and J. R. Westbury, "Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations," *The Journal of the Acoustical Society of America*, vol. 75, no. 3, pp. 945–951, 1984.

[5] J. Šimko and F. Cummins, "Embodied task dynamics," *Psychological Review*, vol. 117, no. 4, pp. 1229—-1246, 2010.

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

[7] A. Windmann, J. Šimko, and P. Wagner, "Optimization-based modeling of speech timing," *Speech Communication*, vol. 74, pp. 76–92, 2015.

[8] J.-F. Patri, J. Diard, and P. Perrier, "Optimal speech motor control and token-to-token variability: a bayesian modeling approach," *Biological Cybernetics*, vol. 109, pp. 611–626, 2015.

[9] B. Parrell and A. C. Lammert, "Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives," *Frontiers in Psychology*, vol. 10, p. 2251, 2019.

[10] A. Turk and S. Shattuck-Hufnagel, *Speech timing: Implications for theories of phonology, speech production, and speech motor control*. Oxford University Press, USA, 2020, vol. 5, ch. How do timing mechanisms work?, pp. 238–263.

[11] ——, "Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production," *Frontiers in Psychology*, vol. 10:2952, 2020.

[12] T. Rakotomalala, P. Perrier, and P. Baraduc, "Trajectories predicted by optimal speech motor control using lstm networks," in *Interspeech 2022-23rd Annual Conference of the International Speech Communication Association*. ISCA; ISCA, 2022, pp. 630–634.

[13] B. Elie, J. Šimko, and A. Turk, "Optimization-based planning of speech articulation using general Tau Theory," *Speech Communication*, vol. 160, p. 103083, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639324000554

[14] M. G. Pandy, F. E. Zajac, E. Sim, and W. S. Levine, "An optimal control model for maximum-height human jumping," *Journal of Biomechanics*, vol. 23, no. 12, pp. 1185–1198, 1990.

[15] G. Schultz and K. Mombaur, "Modeling and optimal control of human-like running," *IEEE/ASME Transactions on Mechatronics*, vol. 15, no. 5, pp. 783–792, 2009.

[16] O. S. Oguz, Z. Zhou, S. Glasauer, and D. Wollherr, "An inverse optimal control approach to explain human arm reaching control based on multiple internal models," *Scientific Reports*, vol. 8, no. 1, p. 5583, 2018.

[17] G. Faity, V. R. Barradas, N. Schweighofer, and D. Mottet, "Force reserve predicts compensation in reaching movement with induced shoulder strength deficit," *Journal of Neurophysiology*, vol. 132, no. 2, pp. 470–484, 2024.

[18] R. Shadmehr, J. J. O. De Xivry, M. Xu-Wilson, and T.-Y. Shih, "Temporal discounting of reward and the cost of time in motor control," *Journal of Neuroscience*, vol. 30, no. 31, pp. 10 507–10 516, 2010.

[19] B. Berret, E. Chiovetto, F. Nori, and T. Pozzo, "Evidence for composite cost functions in arm movement planning: an inverse optimal control approach," *PLoS Computational Biology*, vol. 7, no. 10, p. e1002183, 2011.

[20] K. Friston, "What is optimal about motor control?" *Neuron*, vol. 72, no. 3, pp. 488–498, 2011.

[21] D. McNamee and D. M. Wolpert, "Internal models in biological control," *Annual review of control, robotics, and autonomous systems*, vol. 2, no. 1, pp. 339–364, 2019.

[22] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*. Springer, 1990, pp. 403–439.

[23] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[24] B. Elie, J. Šimko, and A. Turk, "Optimization-based modeling of Lombard speech articulation: Supraglottal characteristics," *JASA Express Letters*, vol. 4, no. 1, p. 015204, 01 2024. [Online]. Available: https://doi.org/10.1121/10.0024364

[25] J. Šimko, M. O'Dell, and M. Vainio, "Emergent consonantal quantity contrast and context-dependence of gestural phasing," *Journal of Phonetics*, vol. 44, pp. 130–151, 2014.

[26] D. P. Bertsekas, "Dynamic programming and optimal control," *Journal of the Operational Research Society*, vol. 47, no. 6, pp. 833–833, 1996.

[27] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[28] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. John Wiley & Sons, 2012.

[29] B. Elie, J. Šimko, and A. Turk, "Optimal control theory of speech production using probabilistic articulatory-acoustic models," in *20th International Conference of Phonetic Sciences (ICPhS)*, 2023.

[30] ——, "Optimal control of speech with context-dependent articulatory targets," in *Interspeech 2023, Dublin*, 2023.

[31] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[32] R. Shadmehr and J. W. Krakauer, "A computational neuroanatomy for motor control," *Experimental Brain Research*, vol. 185, no. 3, pp. 359–381, 2008.

[33] S. H. Scott, "The computational and neural basis of voluntary motor control and planning," *Trends in Cognitive Sciences*, vol. 16, no. 11, pp. 541–549, 2012.

[34] 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, p. e1007321, 2019.

[35] B. Elie, "Planart," https://git.ecdf.ed.ac.uk/belie/planart, 2023.

[36] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, pp. 369–372.

[37] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research*, vol. 12, no. 7, 2011.

[38] B. H. Story and K. Bunton, "Vowel space density as an indicator of speech performance," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. EL458–EL464, 2017.