INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Optimal control of speech with context-dependent articulatory targets

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

[1]Linguistics and English Language; School of Philosophy, Psychology and Language Sciences; the University of Edinburgh; Edinburgh, Scotland, United Kingdom
[2]Faculty of Arts; University of Helsinki; Helsinki, Finland

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

Abstract

This paper presents a computational implementation of phonetic planning which consists of choosing the position of articulatory targets which satisfy conflicting linguistic and extra-linguistic requirements. We present a minimal model that considers intelligibility and least effort as task requirements. To achieve the context-dependent variability of targets, our model approximates intelligibility as a function of target phoneme recognition probability given a vector of articulatory parameters. Preliminary experiments show that our minimal computational model of phonetic planning is able to predict two types of hypoarticulation by adjusting the weight assigned to effort: vowel centralization and stop consonant lenition.

Index Terms: Speech production, Articulatory planning, Optimal Control Theory, Lenition

1. Introduction

Many attempts have been made to build a model of speech production that predicts speech articulatory movements (e.g. [1–4]). One dominant model is Articulatory Phonology/Task dynamics (AP/TD) [1, 5, 6]. AP/TD assumes that speech articulatory movements follow a gestural score which defines the activation time of pre-defined gestures that creates speech sounds. During the activation of a gesture, articulators move such that the timecourse of tract variables (i.e., constriction degree and/or location) tend asymptotically toward a invariant target, following the second-order equation of movements of critically damped oscillators. This approach has been coupled with Optimal Control Theory (OCT) [7, 8] in Embodied Task Dynamics (ETD) [3], where the gestural score and the gestural stiffness are chosen such that they optimize a multi-factor objective function.

Recently, Turk and Shattuck-Huffnagel [4, 9] proposed XT/3C, a phonology-extrinsic-timing, 3-component model of speech production that differs from AP/TD in several key respects. For example, XT/3C has separate Phonological and Phonetic stages. In the Phonological planning stage, task requirements are prioritized and qualitative acoustic cues are chosen for each contrastive element as appropriate to its context. In Phonetic Planning, articulatory trajectories that produce desired acoustic cues are planned according to the equations of Lee's Tau theory [10]. This dynamic articulatory model is an interpolation function that generates the timecourse of movements which close a gap in a given duration. Unlike the task-dynamics model, which uses asymptotic target positions, Tau theory assumes that targets are reached, and that they are reached at an explicitly targeted endpoint time. As a consequence, XT/3C assumes that some of the variation in speech is due to variability in the time and position of targets, as opposed to the undershoot of invariant targets. The choice of these targets is done via an OCT-based optimization.

Since, to the best of our knowledge, no computational implementation of XT/3C has been published yet, this paper presents a first attempt to computationally model XT/3C's Phonetic Planning. This consists of a minimal implementation of the model in order to verify that it can predict basic features of speech. A computational model of XT/3C's Phonetic Planning involves the specification of the objective function used in the optimization process. However, due to the variable nature of targets in XT/3C, the intelligibility cost function cannot be considered as a distance from an invariant canonical target, as assumed in previous AP-based speech optimization models [3, 11]. Instead, we propose an approach which consists of modeling intelligibility as the probability of a target speech sound (e.g. a phoneme) to be recognized given an articulatory configuration. This approach offers a flexible way to account for different languages, language varieties, and speech variation within utterances. This is because probabilistic models can be modified for specific languages as long as labeled corpora are available for each language. Some aspects of speech variation can be modeled within the Phonetic Planning component of XT/3C by varying the weights assigned to each component of the cost function (task requirements and movement costs).

The structure of the paper is as follows. Section 2 introduces a minimal OCT-based model as an initial development of our computational model of XT/3C's Phonetic Planning component. Section 3 introduces the static and articulatory models used for our experiments. Section 4 details the methodology for computing the probabilistic intelligibility model used during the optimization process. Section 5 presents short preliminary experiments to illustrate the interest of the approach: We evaluate the impact of the weight assigned to effort on i) the positions of the optimized vowels in formant space, and ii) the production of consonants in an intervocalic context.

2. Modeling optimal control of speech with context-dependent articulatory targets

This paper presents a minimal working model of speech production using OCT. For that purpose, it considers only intelligibility and least articulatory effort as tasks to satisfy. We assume that this minimal optimization procedure should be sufficient to predict basic features of speech, such as hypo- and hyperarticulation as predicted by Lindblom's H&H theory [12]. This

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; Grant awarded to A. Turk; Grant agreement No. 101019847).

Experiments reported in this paper used the PlanArt software, available at git.ecdf.ed.ac.uk/belie/planart.

10.21437/Interspeech.2023-172

paper does not consider the requirement of utterance brevity, as additionally proposed in other papers [3, 11]; this could be considered in future work. The objective function is thus:

$$C(\theta) = \alpha_E E(\theta) + \alpha_I (1 - I(\theta)), \tag{1}$$

where $C(\theta)$, $E(\theta)$, and $I(\theta)$ are the overall cost, the effort cost, and the intelligibility cost, respectively, all functions of the model parameter vector θ. Maximizing intelligibility is assumed to lead to hyperarticulation, whereas minimizing articulatory effort is assumed to lead to hypoarticulation. These conflicting demands can be balanced and modulated by adjusting the weights α_E and α_I, assigned to the effort and the intelligibility costs, respectively.

2.1. The Effort cost

Two dominant performance objectives are commonly used in Optimal Control Theory (OCT) to quantify effort, namely minimum squared jerk [13–15], and minimum squared motor commands [16–18]. In this paper, we chose to consider the minimum squared motor command objective function. As proposed in [3, 19], we assume that the motor commands are defined as the resulting forces acting on the articulator. Following Newton's second law of motion, the force acting on the nth articulator (or the nth static parameter, as explained in Sec. 3.1) is $F_n(t) = m_n \ddot{x}_n(t)$, where m is the mass of the articulator, and $x_n(t)$ is time course of the nth articulator. This yields the following cost accounting for the effort of articulatory parameter n:

$$E_n = \int_0^T |F_n(t)|^2 dt = m_n^2 \int_0^T |\ddot{x}_n(t)|^2 dt. \tag{2}$$

Since there is no known principled way to tune the mass m for each articulatory parameter, we decided to simply set them all to $m = 10^{-3}$kg. The total effort cost is the sum of the effort associated to each articulator, namely $E = \sum_{n=1}^{N} E_n$.

2.2. The Intelligibility cost

In Embodied Task Dynamics [3], derived from AP/TD, intelligibility is measured as the distance between the current state of the system and a pre-defined invariant target. However, the distance-to-invariant-target approach is inappropriate for XT/3C, because XT/3C's targets are context-dependent. That is, because the targets for the same phoneme in different contexts differ from each other, a different approach is required.

In this paper we propose a novel approach to speech targets and intelligibility based on probabilistic articulatory-acoustic models. The idea is to consider the intelligibility function as an approximation of the probability of recognition of a target speech sound given an articulatory configuration. On the assumption that human perception of phonemes is based on the statistical distributions of their characteristics in the acoustic and/or articulatory space, our approach provides a principled approach to approximate intelligibility during speech communication. In addition, this model can be applied to any type of model, whether the model requires invariant targets or context-dependent targets. In XT/3C, it is assumed that the variability of targets will result from the optimization of these targets, and that speech variation will result from tuning the weights assigned to each task of the global objective function. Mathematically, the intelligibility I is proportional to the posterior probability $P(p|\mathbf{x})$ of the phone p given a static articulatory vector $\mathbf{x}$ (*cf.* Sec. 3.1 for the definition of $\mathbf{x}$). It is defined as:

$$I = \max\left(P(p|\mathbf{x})\right) \times \frac{2}{\pi} \arctan\left(c\Delta t\right), \tag{3}$$

where $P(p|\mathbf{x})$ is the posterior probability of the phone p given the static articulatory vector $\mathbf{x}$. Δt is the time duration of a segment for which $P(p|\mathbf{x})$ is higher than an ad hoc threshold, and c is a constant. The term $\frac{2}{\pi} \arctan(c\Delta t)$ is used to account for the non-linear relationship between phone intelligibility and phone duration. Indeed, the probability of phone recognition increases asymptotically for longer durations [20, 21]. Following the idea by Šimko and Cummins [3], we propose the $\arctan$ function to model the non-linear function of phone probability as a function of time. The constant c was chosen to adjust the shape of the function. High values of c return high phone probabilities for short phone segments, while small values of c require longer phone segments for high phone probabilities.

Section 4 details the methods for implementing probabilistic models, which were used to compute $P(p|\mathbf{x})$. In this paper we set $c = 500$, and Δt as the time segment for which $P(p|\mathbf{x}) \geq \frac{2}{3}\max\left(P(p|\mathbf{x})\right)$. This relative threshold was chosen to ensure a non-null intelligibility gradient, which would prevent the optimization process from converging.

3. Articulatory models

The variable θ (the input of the objective function in Eq. (1)) combines two types of parameters: static and dynamic. The static parameters define a static articulatory model describing the position of the speech articulators and the geometry of the vocal tract at a given time instant. Dynamic parameters pertain to a dynamic articulatory model describing the time-course of the static parameters' movements.

3.1. The static articulatory model

The Maeda model [22] was used as the static articulatory model. This articulatory model generates midsagittal shapes of the vocal tract using seven independent articulatory parameters, corresponding to the principal components that explain most of the observed variance in articulatory data. These are expressed in terms of standard deviations above or below the mean value, where the mean value (i.e. 0) corresponds to a neutral position. The static parameters at a given instant t are stored in the vector $\mathbf{x}$, containing the values of the seven parameters, where each value is contained between -3 and +3.

3.2. The dynamic articulatory model

The dynamic articulatory model used in XT/3C is general Tau theory [10], which states that voluntary movements close a gap between the current state of an effector and its target state. As opposed to asymptotic models used in AP/TD [1, 23, 24], Tau theory assumes that targets are reached. Given an initial gap X_0 and gap-closure duration T, the gap-closing function $X(t)$ depends only on the Tau-coupling parameter k:

$$X(t) = X_0 \left(1 - \frac{t^2}{T^2}\right)^{\frac{1}{k}}. \tag{4}$$

Modifying k will shape the velocity profile, as shown in Figure 1.

4. The intelligibility model

The intelligibility model is trained and tested using a set of synthetic data, denoted $\mathcal{X}$, consisting of 10^7 randomly generated

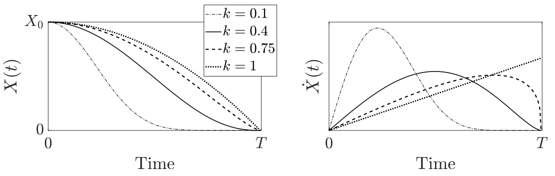

Figure 1: *Examples of gap functions and velocity profiles of Tau-guided movements. The left plot displays Tau-guided movements for different k-values. The right plot displays their corresponding velocity profile.*

vectors containing the values of the 7 parameters of the Maeda model, following a uniform distribution between -3 and 3.

4.1. Labeling the training data

Labeling $\mathcal{X}$ with the appropriate phoneme was done as follows. The first step consisted of identifying stop consonants, fricatives and vowels. We computed tract variables following the technique detailed in [25] and used them to detect and label consonants according to the degree and location of the minimal constriction. The tract variables are lip aperture (LA), tongue tip constriction degree ($TTCD$), and tongue body constriction degree ($TBCD$). In addition, we estimated the minimal constriction degree $\text{MCD} = \min\{LA, TTCD, TBCD\}$ and the location of the minimal constriction $\text{MCL} = \text{argmin}\{LA, TTCD, TBCD\}$. A configuration is considered as a consonant if MCD < 0.5 cm, otherwise it is considered a vowel. A consonant is considered as a stop consonant if MCD < 0.1 cm. Otherwise, the consonant is considered as a fricative. The stop consonants were labelled as /b/, /d/ and /g/ for MCL at lips, tongue tip and tongue body, respectively. The corresponding labels for fricatives were /v/, /z/, and /ʒ/, respectively, if the consonant was a fricative.

Labeling vowel configurations was required to predict the vowel from the formant pattern. For that purpose, we specifically built a Gaussian Mixture Model (GMM) trained on real formant data. Training data were extracted from the Vocal Tract Resonance (VTR) Corpus [26], which contains manually extracted formant trajectories of 538 utterances from the TIMIT database [27], uttered by 186 speakers of American English. The values of the 4 first formants at the mid-point of each of the 5526 analyzed monophthong vowels in the VTR corpus were extracted, resulting in a 5526×4 matrix. We then merged some vowels as follows: /ə/-like vowels `ax`, `axr`, and `ax-h`, merging into a single `ax` class, /ɪ/-like vowels `ix` and `ih` merging into a single `ih` class, and /u/-like vowels `uw` and `ux` merging into a single `u` class. After merging, there were 11 vowel classes. Vocal Tract Length Normalization (VTLN) was applied to formant values [28]. We chose the length of the normalized vocal tract L_{ref} to correspond to the length of the vocal tract in the neutral configuration of the Maeda model, namely $L_{\text{ref}} = 16.27$cm. The GMM was fitted on the data using the iterative Expectation-Maximization algorithm. It was used only to label the vowel configurations in $\mathcal{X}$ with the predicted vowel.

4.2. Training articulatory-to-probability models

We trained a Multi-Layer Perceptron with one hidden layer of 100 nodes as a classifier. We randomly took 90% the data in $\mathcal{X}$ taken at random for training. The remaining data was used as the validation set. We chose a Rectified Linear Unit (ReLU) as activation function and a batch size of 1024 for training. We used a L2 regularization term applied to the model weights in order to allow smoother decision boundaries. A value of 1 was empirically found to be a good trade-off between high classification accuracy and sufficiently smooth decision boundaries. The accuracy score on the validation set was 95.2% .

5. Experiments

We present two short experiments which aim at verifying the ability of our minimal implementation of XT/3C's phonetic planning for predicting speech phenomena. Both experiments investigate the impact of the weight assigned to effort on the production of speech sounds. The first experiment focuses on the production of vowels. The second experiment focuses on the production of stop consonants in an intervocalic context. The balance between effort and intelligibility is quantified by an effort ratio defined as $\frac{\alpha_E}{\alpha_I}$. In these experiments, $\alpha_I = 1$, such that the effort ratio is equivalent to α_E.

5.1. The impact of the least effort requirement on the production of vowels

For each individual optimization of a set of 5 vowels (/ɑ, ɔ, u, i, e/), we consider two movements. The first one goes from the Maeda's neutral position (all static parameters are set to 0) to the vowel to optimize. The second one goes back to the Maeda's neutral position from the optimized vowel target. Each movement duration was 250 ms long. The k-values of the second movement were fixed to 0.4, which corresponds to a purely symmetrical velocity profile. A symmetric velocity profile was chosen because it has been frequently observed in previous studies [29–31]. The parameters to be optimized were the set of the first movement's endpoints for each articulatory parameter, as well as a global k-value corresponding to the first movement, which will be considered as the same for all articulators, i.e., a parameter vector $\theta = [k, x_1, x_2, \ldots, x_7]^T$.

For the 5 vowels of our phone set, we ran several optimization procedures with various effort weights α_E. We varied α_E from 0 to 10000. For each weight value, we ran 50 optimization processes with different initial solutions generated randomly. The final solution was then the one that returned the lowest cost. Figure 2 shows the position of the returned solutions in the vocalic space $F1$–$F2$ for the different weights assigned to articulatory effort. Formant frequencies have been estimated from the vector of Maeda's parameters included in the optimized θ. The right plot shows the k-values returned by the optimization as a function of α_E. As expected, increasing the weight assigned to the effort cost results in vowel centralization: vowel positions in formant space converge towards a central position, corresponding to the formants of the neutral configuration of the Maeda model. As a consequence, the volume of the vocalic space becomes smaller as the effort weight increases. The weight assigned to effort also has an impact on the optimized k-values. A small effort weight leads to small k-values. This is because small k-values correspond to movement with early velocity peaks: they move quickly to a position close to the target. In our model, this increases intelligibility because the movement allows the articulatory configurations to spend more time in a high-probability region, resulting in an increase of Δt in Eq. (3). However, as shown in [31], Tau-guided movements with small k-values require much more effort than the effort-optimized Tau-guided movements. Consequently, this gain of intelligibility with early peak velocity is compensated by the effort cost when α_E increases: optimized k-values increases for $\alpha_E \geq 10^{-3}$ in an asymptotic manner towards the optimized

k-value of 0.45 for $\alpha_E \geq 1$. This k-value corresponds to the value for which a Tau-guided movement will produce the minimal effort for a given amplitude and duration, as shown in [31].

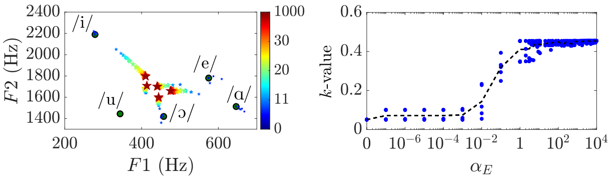

Figure 2: *Left: position of the solutions in the $F1 - F2$ vocalic space for 5 vowels (/ɑ, ɔ, u, i, e/). The size and colors of star markers correspond to different values of α_E. The larger the star marker, the greater the weight. The initial position for each vowel ($\alpha_E = 0$) is denoted by a green circle. Right: optimized k-values as a function of α_E plotted on a logarithmic scale. The dashed line represents the median k-value for a given α_E.*

5.2. The impact of the least effort requirement on the production of stop consonants in an intervocalic context

We consider an articulation consisting of four movements: (1) from the Maeda's neutral position to a vowel configuration (denoted V1), (2) from V1 to the target consonant (C), (3) from C to the second vowel (V2), and (4) back to the neutral position. In this experiment, V1 and V2 are fixed: only the target position C and the global k-value of the consonantal movement are optimized. For the sake of brevity, we only show results for the alveolar stop consonant (C=/d/) with V1=V2=/ɑ/. The articulatory vector used for V1 and V2 is $\mathbf{x}_V = [0.51, 2.52, -1.31, -3, 3, -3, -1.06]^T$. Similarly to the first experiment, eight parameters are optimized, namely $\theta = [k, x_1, x_2, \ldots, x_7]^T$.

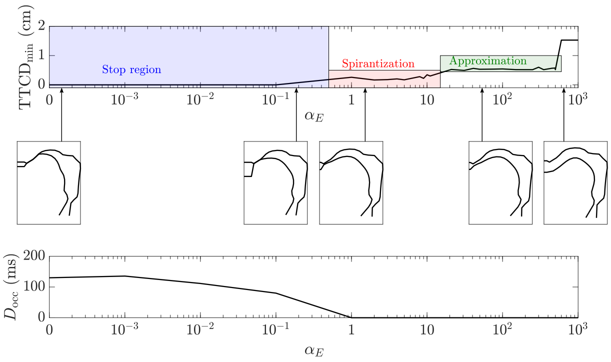

Figure 3: *Minimum of the tongue tip constriction degree (Top) and duration of occlusion (bottom) in the optimized /ɑdɑ/ sequence, as a function of the weight assigned to effort (α_E) in a logarithmic scale. The center of the plot displays the contours of the vocal tract corresponding to the optimized consonant targets, following the Maeda model.*

Fig. 3 shows the minimal constriction of the tongue tip constriction degree and the duration of the alveolar occlusion during the simulated /ɑdɑ/ sequence as a function of α_E, the weight assigned to effort. The evolution of the minimal tongue tip constriction degree (top panel of Fig. 3) shows different regions. For small α_E (≤ 1), the consonantal movement has a sufficiently large amplitude to allow the formation of an occlusion (the minimal $TTCD$ is 0): this is the stop region. Note that the linguopalatal contact region may change in this region (see the vocal tract shapes in the middle panel of Fig. 3). The contact region is much larger for very small α_E compared to α_E close to 1. As seen in the bottom plot of Fig. 3, the occlusion duration increases with decreasing α_E.

When $\alpha_E \geq 1$, the optimal solution does not allow the vocal tract to be closed at the tongue tip. The minimal $TTCD$ is larger than 0, hence no occlusion. Depending on the value of the minimal $TTCD$, three regions can be identified: spirantization, approximation, and vocalization. Spirantization occurs when $1 \leq \alpha_E \leq 20$: the minimal $TTCD$ is smaller than 0.5 cm, which corresponds to a fricative. When $30 \leq \alpha_E \leq 500$, this is the approximation region: the minimal $TTCD$ is below 1 cm and above 0.5 cm, which yields an approximant. Finally, when $\alpha_E > 500$, the effort requirement is too large to allow a movement towards a consonant: the returned solution is a vowel similar to V1 and V2.

6. Conclusion and future work

This paper presents a simple optimization-based computational model of XT/3C's phonetic planning [4, 9]. It used a minimal model that accounts only for effort *vs.* intelligibility during the multi-task optimization process. The paper presents two preliminary experiments which show that our computational implementation of XT/3C's phonetic planning is able to predict some basic features of speech: vowel centralization and stop consonant lenition in hypoarticulated speech. These results provide support for the use of XT/3C as an articulatory planning model. In the future, the objective function should be developed to account for different types of timing effects, such as timing patterns relating to prosodic structure [11] and rate of speech. All experiments reported in this paper used the PlanArt software, publicly available at `git.ecdf.ed.ac.uk/belie/planart`.

It is of note that our model predicted the use of two distinct regimes of dynamic articulatory trajectories: an early peak velocity regime and a nearly symmetric velocity profile. The early peak velocity regime is predicted when the least effort requirement is small in relation to the intelligibility requirement, which allows the speaker to spend more time in an intelligible, high probability region. When the least effort requirement is sufficiently large, the nearly symmetric velocity profile regime is favored as it requires much less effort than the early peak velocity regime [31]. Previous experimental studies showed that speakers primarily use a nearly symmetrical velocity profile [29–31], which supports the hypothesis of a least effort requirement during speech production.

This paper also presents a new probabilistic approach to account for intelligibility in OCT-based models of speech production. As the model is trained partially on real speech data, it provides a realistic approximation of intelligibility that relates intelligibility to observed distributions of articulatory characteristics in a principled way. Although this probability model has been designed for our context-dependent articulatory targets, it can also be applied to speech production models that consider invariant targets. In addition, probabilistic models can be modified for a specific language or variety of language if an appropriate labeled corpus is available. One possible improvement would be to build intelligibility models solely based on real data, using an articulatory-acoustic database. This would also allow more phonemes to be considered and more realistic aspects of speech to be simulated.

7. References

[1] E. Saltzman, “Task dynamic coordination of the speech articulators: A preliminary model,” in *Generation and Modulation of Action Patterns (Experimental Brain Research Series)*, vol. 15. Springer Berlin, Heidelberg, Berlin, 1986, pp. 129–144.

[2] F. H. Guenther, “Neural control of speech movements,” in *Phonetics and phonology in language comprehension and production: Differences and similarities*, N. O. Schiller and A. S. Meyer, Eds. Mouton de Gruyter, Berlin, 2003, pp. 209–239.

[3] J. Simko and F. Cummins, “Embodied task dynamics,” *Psychological review*, vol. 117, no. 4, pp. 1229—-1246, 2010.

[4] A. Turk and S. Shattuck-Hufnagel, “Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production,” *Frontiers in Psychology*, vol. 10:2952, 2020.

[5] C. P. Browman and L. M. Goldstein, “Towards an articulatory phonology,” *Phonology*, vol. 3, pp. 219–252, 1986.

[6] C. P. Browman, L. Goldstein *et al.*, “Dynamics and articulatory phonology,” in *Mind as motion: Explorations in the dynamics of cognition*, R. F. Port and T. van Gelder, Eds. MIT press Cambridge, MA, 1995, pp. 175–194.

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

[8] R. Shadmehr and J. W. Krakauer, “A computational neuroanatomy for motor control,” *Experimental brain research*, vol. 185, no. 3, pp. 359–381, 2008.

[9] 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.

[10] D. N. Lee, “Guiding movement by coupling taus,” *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

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

[12] B. Lindblom, “Explaining phonetic variation: A sketch of the H&H theory,” in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.

[13] T. Flash and N. Hogan, “The coordination of arm movements: an experimentally confirmed mathematical model,” *Journal of neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[14] B. Hoff and M. A. Arbib, “Models of trajectory formation and temporal interaction of reach and grasp,” *Journal of motor behavior*, vol. 25, no. 3, pp. 175–192, 1993.

[15] D. Sha, J. L. Patton, and F. A. Mussa-Ivaldi, “Minimum jerk reaching movements of human arm with mechanical constraints at endpoint.” *Int. J. Comput. Syst. Signals*, vol. 7, no. 1, pp. 41–50, 2006.

[16] A. H. Fagg, A. Shah, and A. G. Barto, “A computational model of muscle recruitment for wrist movements,” *Journal of Neurophysiology*, vol. 88, no. 6, pp. 3348–3358, 2002.

[17] I. O’Sullivan, E. Burdet, and J. Diedrichsen, “Dissociating variability and effort as determinants of coordination,” *PLoS computational biology*, vol. 5, no. 4, p. e1000345, 2009.

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

[20] W. A. Grimm, “Perception of segments of english-spoken consonant-vowel syllables,” *The Journal of the Acoustical Society of America*, vol. 40, no. 6, pp. 1454–1461, 1966.

[21] M. E. Tekieli and W. L. Cullinan, “The perception of temporally segmented vowels and consonant-vowel syllables,” *Journal of Speech, Language, and Hearing Research*, vol. 22, no. 1, pp. 103–121, 1979.

[22] 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.

[23] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, “A gesture-based dynamic model describing articulatory movement data,” *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1878–1889, 1995.

[24] T. Sorensen and A. Gafos, “The gesture as an autonomous nonlinear dynamical system,” *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[25] J. L. Gaines, K. S. Kim, B. Parrell, V. Ramanarayanan, S. S. Nagarajan, and J. F. Houde, “Discrete constriction locations describe a comprehensive range of vocal tract shapes in the Maeda model,” *JASA express letters*, vol. 1, no. 12, p. 124402, 2021.

[26] 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.

[27] L. F. Lamel, R. H. Kassel, and S. Seneff, “Speech database development: Design and analysis of the acoustic-phonetic corpus,” in *Speech Input/Output Assessment and Speech Databases*, 1989, pp. 161–170.

[28] K. Johnson, “Vocal tract length normalization,” *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[29] D. J. Ostry, J. D. Cooke, and K. G. Munhall, “Velocity curves of human arm and speech movements,” *Experimental Brain Research*, vol. 68, no. 1, pp. 37–46, 1987.

[30] J. S. Perkell and M. Zandipour, “Economy of effort in different speaking conditions. II. Kinematic performance spaces for cyclical and speech movements,” *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1642–1651, 2002.

[31] B. Elie, D. N. Lee, and A. Turk, “Modeling trajectories of human speech articulators using general tau theory,” *Speech Communication*, vol. 151, pp. 24–38, 2023.