Simulating articulatory trajectories with phonological feature interpolation

Angelo Ortiz Tandazo, Thomas Schatz, Thomas Hueber, Emmanuel Dupoux

Interspeech 2024
1-5 September 2024, Kos, Greece

Simulating articulatory trajectories with phonological feature interpolation

Angelo Ortiz Tandazo[1,2], *Thomas Schatz*[3], *Thomas Hueber*[2], *Emmanuel Dupoux*[1,4]

[1]ENS, PSL Research University, EHESS, CNRS, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
[3]Aix-Marseille Univ., CNRS, LIS, France
[4]Meta AI

angelo.ortiz.tandazo@ens.psl.eu, thomas.schatz@univ-amu.fr,
thomas.hueber@grenoble-inp.fr, emmanuel.dupoux@gmail.com

Abstract

As a first step towards a complete computational model of speech learning involving perception-production loops, we investigate the forward mapping between pseudo-motor commands and articulatory trajectories. Two phonological feature sets, based respectively on generative and articulatory phonology, are used to encode a phonetic target sequence. Different interpolation techniques are compared to generate smooth trajectories in these feature spaces, with a potential optimisation of the target value and timing to capture co-articulation effects. We report the Pearson correlation between a linear projection of the generated trajectories and articulatory data derived from a multi-speaker dataset of electromagnetic articulography (EMA) recordings. A correlation of 0.67 is obtained with an extended feature set based on generative phonology and a linear interpolation technique. We discuss the implications of our results for our understanding of the dynamics of biological motion.

Index Terms: speech production, computational modelling, phonological features, articulatory-to-acoustic mapping

1. Introduction

Recent advances in self-supervised learning (SSL) have led to progress in various speech processing tasks [1, 2] and language modelling from speech units [3]. These SSL models require increasingly greater amounts of (unlabelled) data to capture as much acoustic variance as possible. Moreover, their capacity to learn high-level language representations hinges on the quality of the underlying speech units [4], which are not linguistically interpretable [2, 5]. Importantly, these representations remain sensitive to contextual effects such as co-articulation [6] making them sub-optimal to efficiently code context-invariant phonological units.

According to the motor [7] or perceptuo-motor theories [8] of speech perception, humans' 'quest for invariance' [9] is done by recovering motor or articulatory representations from an auditory input. These representations are supposed to be less variable than their acoustic counterparts. Several studies have found neuro-physiological correlates of this mental 'sensory-to-motor' inverse mapping in speech perception [10]. Incorporating such motor representation into SSL speech models could potentially improve their performance (*e.g.* noise robustness, low-resource downstream tasks, etc.) and lead to more plausible computational models of speech and language acquisition.

In a simplified perception-production loop of speech motor control [11] (see the bottom left of Figure 1), the motor commands are derived from the sensory, acoustic signal and used to generate the underlying articulatory trajectories (via a so-called *forward model*).

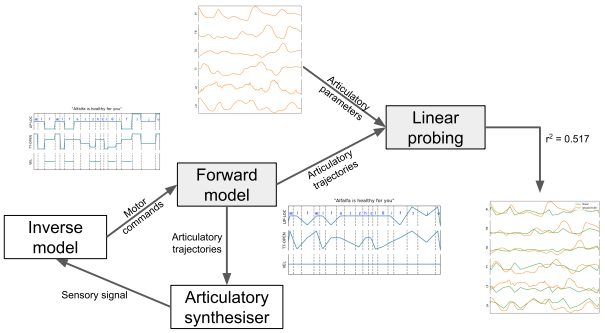

Figure 1: *Simplified diagram of a speech perception-production loop (to the left). The focus of this work lies in the forward model and the linear probing (to the right).*

Several computational models of such speech perception-production loop have been proposed in the literature [12, 13, 14, 15]. However, in most studies, the motor or articulatory representations are derived from a specific speaker or a specific articulatory model. As a result, these models, while enabling to study some of the underlying processes of speech acquisition, perception and motor control by simulation, are not designed to scale up to large numbers of speakers or languages.

As a first step towards a more universal SSL speech model integrating motor or articulatory representations, we focus in the present study on the forward model, *i.e.* from the motor commands to the generation of articulatory trajectories. First, we investigate different feature sets to encode a given phonetic target sequence, by relying either on the generative phonology (GP, [16]) or on the articulatory phonology (AP, [17]) theories. Importantly, a phonetic target is here encoded in terms of phonologically-motivated and articulatory-related categories (*e.g.* the place of articulation for a consonant in GP, the location and degree of a constriction in AP).

To generate continuous (and smooth) trajectories in these feature spaces (with a forward model), we test different interpolation methods, which differ from the dynamic properties desired at each phonetic target (*e.g.* zero and/or continuous velocity at each target). To account for the uncertainty of our timing heuristic and the (potential) target undershoot phenomenon, we also consider variants performing target optimisation both in space (find the offset from the ideal feature value, *e.g.* lips only partly closed) and time (reach a target sooner or later). Moreover, the proposed approach can deal with unspecified features for which the value depends on the context (for instance, the position of a constriction modulated by the vocalic context).

10.21437/Interspeech.2024-2192

The generated trajectories in the GP or AP feature spaces are evaluated using a linear probing technique. A linear model is learnt between the generated trajectories on the one hand, and the parameters of an articulatory model [18] built from a multi-speaker electromagnetic articulography (EMA) dataset on the other. Such an evaluation was also used in [19] to probe HuBERT representations, among other SSL models.

The main contributions and findings of the paper are the following: (i) we propose a general methodology to probe pseudo-motor commands and forward models in a computational model of a speech perception-production loop; (ii) we show that features derived from generative phonology (GP) correlate better with real articulatory recordings than articulatory phonology (AP) ones; (iii) a bit surprisingly, we show that a linear interpolation between these features better captures the dynamics of real articulatory data compared to a more complex one (spline based), with constraints on the velocity and/or acceleration at each phonetic target; (iv) we show that the use of unspecified (context-dependent) phonological features improves performance, probably by allowing the forward model to better account for natural co-articulation patterns.

The code with the data processing and interpolation methods can be found at `https://github.com/angelo-ortiz/articulatory-probing`.

2. Methodology

2.1. Phonological feature set

Two phonological feature sets were used. The first one, based on generative phonology and referred to as the GP feature set, was proposed by [20, Chapter 4]. It describes phonetic targets in terms of 26 manner, laryngeal and place features. The GP feature set is considered under two variants: with unknown support and binary. Following [20, p. 91], some phonemes have zero-value features, notably because their values depend on their local context within an utterance or simply because they are irrelevant to the underlying phoneme. Hence, the ternary-valued (including the zero values) GP feature set is considered as is (to be used by interpolating methods handling unknown values) but also in a binary form, in which a zero value is considered as being the 'absence' of the given feature (thus, negatively valued).

The second feature set is based on articulatory phonology (AP), and the location and degrees of constriction of 5 major articulators in the vocal tract. AP-based features have been successfully used in automatic speech recognition, first within a Bayesian framework [21] to deal with pronunciation variability in spontaneous speech, and then in a DNN-based system [22] to increase the robustness to noise. The AP feature values come in the form of categorical distributions over totally ordered categories [21, p. 126]. Feature values are typically Dirac distributions, except for some phoneme features that depend on the phonemic context. Similarly to the GP feature set, we have an AP unknown variant by considering the non-Dirac distributed feature values as *unknown* values to be found contextually. This feature set is then used in a scalar version, in which each feature-value category is mapped to a real value (scalar AP: 8 features); and a one-hot version, in which the feature-value categories are one-hot encoded (one-hot AP: 32 features).

To ensure that the feature-level information for phonemes is relevant, we also use a *feature* set with one-hot phoneme encodings. In total, we evaluate seven feature sets: one-hot phonemes, scalar AP (also enriched with one-hot phonemes), one-hot AP (also enriched with one-hot phonemes), binary GP and unknown-supporting GP (also enriched with one-hot phonemes).

[1] British long vowels and the silence are included.

Table 1: *Articulatory score: average Pearson correlation coefficients of the* 6 *articulatory parameters and* 6 *speakers. The scores correspond to each interpolation method's best configuration. Default: binary features without optimisation. Variants: unknown features*$^{\mu}$*, timing optimisation*†*, timing and position optimisation*‡*. (Standard error across the 6 different speakers was found to be* 0.01 *on average.)*

Feature set	# Features	Method	Score ↑
GP + one-hot phoneme	73	piecewise-cst	0.595
		linear$^{\mu}$	**0.679**
		cubic Hermite$^{\mu}$	0.668
		natural cubic‡	0.663
one-hot AP + one-hot phoneme	94	linear$^{\mu}$	0.663
		cubic Hermite$^{\mu}$	0.648
		natural cubic$^{\mu}$	0.628
scalar AP + one-hot phoneme	70	linear$^{\mu}$	0.656
		cubic Hermite$^{\mu}$	0.642
		natural cubic$^{\mu}$	0.624
one-hot phoneme	47[1]	piecewise-cst	0.589
		linear	0.645
		cubic Hermite‡	0.642
		natural cubic‡	0.630
GP	26	piecewise-cst	0.559
		linear$^{\mu}$	0.630
		cubic Hermite$^{\mu(\dagger)}$	0.622
		natural cubic‡	0.629
one-hot AP	32	linear$^{\mu}$	0.608
		cubic Hermite$^{\mu\ddagger}$	0.596
		natural cubic$^{\mu}$	0.538
scalar AP	8	linear$^{\mu}$	0.511
		cubic Hermite$^{\mu\ddagger}$	0.506
		natural cubic$^{\mu}$	0.479

2.2. Dataset

The articulatory data comes from the publicly available MOCHA-TIMIT dataset[2]. It provides electromagnetic articulography (EMA) recordings for 460 short sentences read by 8 British English speakers along 12 dimensions (2D midsagittal coordinates for 6 articulators: tongue tip, tongue body, tongue dorsum, lower incisor, upper lip and lower lip.) In this study, we consider 6 speakers, namely `fsew0`, `msak0`, `ffes0`, `mjjn0`, `faet0` and `maps0`, because a sequence of waveform files did not match the given transcriptions for the other two speakers. For each speaker, the 460 utterances are split into 410 for training (out of which 20 are randomly drawn for development) and 50 for testing.

The EMA data is low-pass filtered at 50 Hz and downsampled from 500 Hz to 100 Hz. As with [23], raw EMA data is then converted into an easier-to-interpret and lower-dimensional set of 6 'articulatory parameters' (jaw height,

[2] The dataset can be found at `https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`

tongue body, tongue back, tongue tip, lip protrusion and lip height) using a linear decomposition technique, which is often referred to as 'guided PCA'. It aims to decouple the jaw from tongue and lip movements and extract independent degrees of freedom from the vocal tract.

Each audio recording was segmented at the phonetic level using the Montreal Forced Aligner[3]. Based on the resulting phonetic segmentation, the initial and final utterance silences were removed and the utterances with non-silence boundary phones were discarded. The filtered phonetic segmentations were later mapped into *featural* segmentations by replacing the phones with the phonological features of their underlying phonemes (lookup table mapping). Finally, we inferred timings for the phonological targets from the time midpoint of each phoneme in the featural segmentation.

2.3. Forward model

Different forward mapping techniques are tested to generate continuous trajectories from the discrete pseudo-motor commands provided by GP and AP (Section 2.1). As a first baseline, we use the piecewise-constant interpolation, which keeps all the phonological features constant for the duration given by the phonetic segmentation.

To test a smoothness degree that better fits the articulatory space, we test linear and cubic interpolation methods. Specifically, we consider two cubic methods: the cubic Hermite spline and the natural cubic spline. The former enforces zero velocity at all targets, and continuity of both position and velocity (so the acceleration is possibly discontinuous at the targets); whereas the latter enforces continuity of all position, velocity and acceleration, with zero acceleration at the initial and final targets.

Formally, let $d \in \mathbf{N}_{>0}$ be the number of phonological features. For a given utterance, let $K \in \mathbf{N}_{>0}$ be its number of non-boundary (or intermediate) targets. Then, its featural segmentation is denoted by $(\mathbf{X}, \mathbf{Y}) \in \mathbf{R}^{(K+2)\times d} \times \mathbf{R}^{(K+2)\times 2}$, where $\mathbf{X}$ contains the $K+2$ target positions, and $\mathbf{Y}$ the targets' time intervals. In this work, we remove the boundary silences, so $y_{1,*} = \mathbf{0}_2$, $y_{K+2,*} = y_{K+1,2}\mathbf{1}_2$, and $x_{1,*} = x_{K+2,*} = \mathbf{0}_d$. From this, we deduce a vector of midpoint target timings $\mathbf{t} \triangleq \frac{1}{2} Y \mathbf{1}_2$. The (base) interpolating function for the utterance $(\mathbf{X}, \mathbf{Y})$, expressed as $f(\tau; \mathbf{X}, \mathbf{t}), 0 \leq \tau \leq t_{K+2}$, thus satisfies

$$f(t_k; \mathbf{X}, \mathbf{t}) = x_{k,*}, \tag{1}$$

for all $1 \leq k \leq K+2$.

To tackle the uncertainty of the midpoint-timing heuristic assumed for the base interpolating functions and to allow for target undershoot[4], we include two additive optimisations over time and space. The optimised interpolating function f learns the target positions and timings $(\mathbf{X}', \mathbf{t}')$ such that

$$f(\tau; \mathbf{X}, \mathbf{t}) = g(\tau, \mathbf{X}', \mathbf{t}'), 0 \leq \tau \leq t_{K+2}, \tag{2}$$

where the base interpolating function g satisfies Equation 1, by minimising the objective function

$$L_\lambda(\mathbf{X}', \mathbf{t}') \triangleq \int_0^{t_{K+2}} \|g''(\tau; \mathbf{X}', \mathbf{t}')\|_2^2 \, d\tau + \lambda \sum_{k=2}^{K+1} \|g(t'_k; \mathbf{X}', \mathbf{t}') - x_{k,*}\|_2^2. \tag{3}$$

[3]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[4]Target undershoot occurs when there is not enough time for the forward model to reach some targets.

We optimise the objective function L_λ by on-utterance gradient descent with the initialisation $(\mathbf{X}', \mathbf{t}') = (\mathbf{X}, \mathbf{t})$.

2.4. Linear probing

To evaluate the different interpolation methods described above, we adopted the same metric as used in [19], referred to as the *articulatory score*.

Let $S_i = \{(\mathbf{X}_j, \mathbf{t}_j, \mathbf{Z}_j)_j\}$ be the set of utterances for the ith speaker, with $\mathbf{Z}_j \in \mathbf{R}^{n_{\mathbf{Z}_j} \times 6}$ being the $n_{\mathbf{Z}_j}$ articulatory parameters of the jth utterance. Then, for each interpolation method f and speaker S_i, we learn a linear transformation h_i that minimises the reconstruction loss from the interpolated articulatory trajectories and the expected articulatory parameters as follows

$$\mathcal{L}_i = \frac{1}{|S_i|} \sum_{(\mathbf{X},\mathbf{t},\mathbf{Z}) \in S_i} \frac{1}{n_{\mathbf{Z}}} \sum_{k=1}^{n_{\mathbf{Z}}} \left\| h_i\left(f\left(\frac{k}{100}; \mathbf{X}, \mathbf{t}\right)\right) - z_{k,*} \right\|_2^2. \tag{4}$$

This is done via gradient descent with a learning rate of 0.001 via the Adam optimiser, with $\beta = (0.9, 0.999)$. We run the learning procedure for 100 epochs unless it is early stopped when the validation loss stops decreasing (patience fixed at 5).

For the optimised cubic interpolations, we first do a grid search of the hyper-parameters used in the on-utterance target optimisation: (i) timing learning rate within $\{10^{-6}, 5\times10^{-6}, 10^{-5}, 5\times10^{-5}, 10^{-4}\}$, (ii) position learning rate within $\{10^{-3}, 10^{-2}, 10^{-1}\}$, and (iii) loss weight parameter $\lambda \in \{0, 10^3, 10^4, 10^5, 10^6, 10^7\}$[5]. For each interpolation method, the hyper-parameter configuration with the highest articulatory score in the development set was selected. Finally, we compute the Pearson correlation coefficient (PCC) between the learnt linear projections and the articulatory parameters. The final articulatory score of an interpolation method is then the average PCC of all articulatory parameters and speakers.

3. Results

We run the linear, cubic Hermite spline and natural cubic spline interpolation methods (Section 2.3) on the seven feature sets derived from one-hot phoneme encoding, AP and GP theories (Section 2.1). The piecewise-constant interpolation can only be run on fully specified feature sets, here the one-hot phoneme and binary GP feature sets.

Table 1 reports the articulatory score by feature space and interpolation method on a held-out test. The scores correspond to each interpolation method's best configuration on each feature set. We observe that the scalar AP proves to be (very) difficult to interpolate on, but replacing the fixed values (probably needing learning) with equidistant values in the form of one-hot encodings helps close the gap to the GP features. Surprisingly, the one-hot phoneme encodings are the best *single* feature set.

Given the potential complementarity of information between the GP/AP feature sets and the one-hot phoneme encodings, we probe the GP and AP features enriched with the latter. The mix turns out to be beneficial for all the interpolation methods, although we lose the reduced number of interpretable features sought for the inverse models in perspective.

Tables 2 and 3 show the articulatory scores per speaker and articulatory parameter, respectively, on the best feature space, namely GP features enriched with one-hot phonemes. In both

[5]The spatial term of the loss in Equation 3 is very small compared with the smoothness term.

Table 2: *Articulatory score for each speaker on the GP feature set enriched with one-hot phonemes.*

Method	msak0	fsew0	ffes0	mjjn0	maps0	faet0	Average
piecewise-cst	0.634	0.616	0.590	0.591	0.537	0.604	0.595
linear	**0.729**	**0.704**	**0.666**	**0.658**	**0.623**	**0.693**	**0.679**
cubic Hermite	0.718	0.695	0.655	0.649	0.611	0.681	0.668
natural cubic	0.711	0.686	0.652	0.632	0.613	0.685	0.663

Table 3: *Articulatory score for each articulatory parameter on the GP feature set enriched with one-hot phonemes.*

Method	Jaw height	Tongue body	Tongue dorsum	Tongue tip	Lip protrusion	Lip height	Average
piecewise-cst	0.646	0.653	0.543	0.539	0.532	0.658	0.595
linear	**0.715**	**0.750**	**0.625**	**0.627**	**0.621**	**0.736**	**0.679**
cubic Hermite	0.703	0.742	0.618	0.616	0.610	0.722	0.668
natural cubic	0.708	0.733	0.604	0.606	0.615	0.713	0.663

cases, the ranking induced by the average score (linear $\succ$ cubic Hermite $\succ$ natural cubic $\succ$ piecewise constant) is met throughout the conditions, bar the two speakers maps0 and faet0, the jaw height and the lip protrusion (natural cubic $\succ$ cubic Hermite). Interestingly, from Tables 1, 2 and 3, it is clear that the linear interpolation method better exploits the given phonological spaces, regardless of the feature nature, speaker or articulatory parameter.

In Table 1, we see that most of the best scores reported were obtained on features with unknown support. The results in Table 4 support the hypothesis that, in general, keeping and interpolating unknown feature values is better than associating them with a fixed value. On the other hand, the effect of the target timing and/or position optimisation depends on the interpolation method. For instance, when we optimise the timings on the cubic Hermite spline, the articulatory score does not improve, and the spatial optimisation has the same (negative) impact. This is why we see few cubic Hermite interpolations with target optimisations in Table 1.

Table 4: *Comparison of GP + one-hot phoneme feature set variants and the effect of target optimisation. The two leftmost scores correspond to the binary and unknown-supporting feature-set variants without optimisation, and the right scores to the timing-only and the timing-and-position optimisations on the underlined feature sets.*

Method	Non-optimised		Optimised	
	Binary	Unknown	Time	Time & space
linear	0.659	**0.679**		
cubic Hermite	0.645	<u>**0.668**</u>	0.660	0.649
natural cubic	<u>0.623</u>	0.638	0.624	**0.663**

4. Conclusion

In this study, we have analysed phonological features as potential pseudo-motor commands in a computational model of a speech perception-production loop. We found that: (i) smooth trajectories on generative phonology features correlate better with articulatory parameters than those on articulatory phonology ones, with a correlation coefficient of 0.67 when GP features are enriched with one-hot phoneme encodings, (ii) a linear forward model better captures the dynamics of real articulatory data, but target optimisation (in terms of timing and/or position) helps a cubic model to reduce the gap, (iii) with the AP features, a better correlation coefficient is obtained with a one-hot encoding, in which all the values for a given feature are equidistant to one another, rather than with a fixed, scalar continuous one, (iv) interpolating unknown (or context-dependent) features is better than associating a fixed value with them.

Since interpolating under-specified dimensions of articulatory targets appears to lead to a better fit, future work could try to push this strategy further by incorporating more under-specified dimensions in featural segmentations, thus enabling the smoothness of the forward model's trajectories to better model co-articulation.

Further work should also investigate the reason why linear interpolation of articulatory targets outperforms smoother cubic spline interpolation. This is surprising since articulatory trajectories are not linear, cubic splines are excellent interpolators and the biological motion literature suggests that smooth trajectories should fit articulatory data well [24]. A possible interpretation is that unwarranted assumptions in our analyses cause the observed advantage of linear interpolation methods. Based on our results, the advantage of linear interpolation does not appear sensitive to assumptions regarding target definition (generative vs articulatory, specification, timing, position). It may be the case, however, that the assumption of a fixed inventory of targets at the phonemic level is too optimistic, even for a single speaker, at least without controlling further variables that may modulate target parameters, such as prosodic effects. To test this hypothesis, simulated trajectories could be used to determine if ignoring prosodic effects would predict an advantage for linear interpolation even when using a cubic spline model for the dynamics. Alternatively, our results may indicate that classical results on biological motion, obtained in highly controlled settings, do not accurately characterize the dynamics of biological motion in less restricted environments. To test this hypothesis, our methodology could be applied to more controlled trajectories (e.g. isolated syllables), where it should find that smoother trajectories provide a better fit than linear interpolation. This would validate our methodology, which could then be leveraged to better understand the dynamics of biological motion in the wild.

5. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014739 made by GENCI.

6. References

[1] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[2] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the Zero Resource Speech Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[3] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "AudioLM: a Language Modeling Approach to Audio Generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[4] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[5] M. Lavechin, M. de Seyssel, M. Métais, F. Metze, A. Mohamed, H. Bredin, E. Dupoux, and A. Cristia, "Modeling early phonetic acquisition from child-centered audio data," *Cognition*, vol. 245, p. 105734, 2024.

[6] M. Hallap, E. Dupoux, and E. Dunbar, "Evaluating context-invariance in unsupervised speech representations," in *Proc. Interspeech*, 2023, pp. 2973–2977.

[7] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, p. 1–36, Oct 1985.

[8] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.

[9] J. S. Perkell and D. H. Klatt, "Invariance and variability in speech processes." in *Symposium on Invariance and Variability of Speech Processes, Oct, 1983, Massachusetts Inst. of Technology, Cambridge, MA, US*. Lawrence Erlbaum Associates, Inc, 1986.

[10] J. I. Skipper, J. T. Devlin, and D. R. Lametti, "The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception," *Brain and Language*, vol. 164, pp. 77–105, 2017.

[11] M. I. Jordan and D. M. Wolpert, "Computational motor control," M. Gazzaniga, Ed. MIT Press, 1999.

[12] 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, no. 6, pp. 611–626, 2015.

[13] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *Proc. International Conference on Development and Learning and on Epigenetic Robotics*, 2014, pp. 195–200.

[14] H. Rasilo and O. Räsänen, "An online model for vowel imitation learning," *Speech Communication*, vol. 86, pp. 1–23, 2017.

[15] M.-A. Georges, J. Diard, L. Girin, J.-L. Schwartz, and T. Hueber, "Repeat after me: Self-supervised learning of acoustic-to-articulatory mapping by vocal imitation," in *Proc. ICASSP*, 2022, pp. 8252–8256.

[16] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper and Row, 1968.

[17] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[18] 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*, W. J. Hardcastle and A. Marchal, Eds. Springer Science & Business Media, 2012, vol. 55, pp. 131–149.

[19] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *Proc. ICASSP*, 2023, pp. 1–5.

[20] B. Hayes, *Introductory Phonology*. John Wiley & Sons, 2008, vol. 7.

[21] K. Livescu, "Feature-based pronunciation modeling for automatic speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[22] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," *Computer Speech Language*, vol. 36, pp. 173–195, 2016.

[23] A. Serrurier, P. Badin, A. Barney, L.-J. Boë, and C. Savariaux, "The tongue in speech and feeding: Comparative articulatory modelling," *Journal of Phonetics*, vol. 40, no. 6, pp. 745–763, 2012.

[24] 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.