Interspeech 2024
1-5 September 2024, Kos, Greece

Towards a Quantitative Analysis of Coarticulation with a Phoneme-to-Articulatory Model

Chaofei Fan[1], *Jaimie M. Henderson*[1], *Chris Manning*[1], *Francis R. Willett*[1,2]

[1]Stanford University, USA
[2]Howard Hughes Medical Institute at Stanford University, USA

stfan@stanford.edu

Abstract

Prior coarticulation studies focus mainly on limited phonemic sequences and specific articulators, providing only approximate descriptions of the temporal extent and magnitude of coarticulation. This paper is an initial attempt to comprehensively investigate coarticulation. We leverage existing Electromagnetic Articulography (EMA) datasets to develop and train a phoneme-to-articulatory (P2A) model that can generate realistic EMA for novel phoneme sequences and replicate known coarticulation patterns. We use model-generated EMA on 9K minimal word pairs to analyze coarticulation magnitude and extent up to eight phonemes from the coarticulation trigger, and compare coarticulation resistance across different consonants. Our findings align with earlier studies and suggest a longer-range coarticulation effect than previously found. This model-based approach can potentially compare coarticulation between adults and children and across languages, offering new insights into speech production.

Index Terms: coarticulation, speech production, phoneme-to-articulatory model

1. Introduction

Coarticulation is the phenomenon in which one phoneme influences the acoustic and articulatory characteristics of its neighbors. Understanding coarticulation is crucial for understanding the mechanisms of speech production, the planning and coordination of speech articulators, and how these processes relate to speech disorders and speech acquisition in children [1].

The analysis of coarticulation has primarily relied on acoustic and articulatory recordings. However, the challenges in data collection have often limited these analyses to simple phonemic sequences (e.g., consonant-vowel, vowel-consonant-vowel), a narrow selection of phonemes, and specific articulatory movements or spectral formant comparisons [2, 3, 4]. Questions remain about the full temporal scope of coarticulation across all phonemes and how its degree and impact might differ among speakers and across languages [5, 6].

One potential solution is to use a speech production model to generate accurate synthetic data for any desired experimental condition. Formal speech production models, such as DIVA [7] and TADA [8], are built on years of observation but are primarily used to produce articulatory kinematics for short phonemic sequences. Recent data-driven phoneme-to-articulatory (P2A) and acoustic-to-articulatory (A2A) models [9, 10] can accurately produce EMA trajectories for long sentences, but they do not provide the timing of phonemes.

Our first contribution is to develop a data-driven P2A model that can make realistic predictions for what an "average" EMA time series would be for any given phoneme sequence, while also returning the time at which each phoneme occurs. With this model, we can sample EMA trajectories for any word and then interrogate the model outputs for any coarticulation effect of interest. Our second contribution is a metric for quantifying the temporal extent and magnitude of coarticulation across all phonemes and articulators. We use minimal word pairs (two words that differ in only one phonological element) to measure how coarticulation spreads from the differing phoneme (trigger) to its neighboring phonemes (targets). Compared to previous studies quantifying coarticulation [11, 12, 13, 14, 15], ours samples a large number of phonemic sequences and leverages the statistical power of 6+ hours of EMA training data from 12 speakers incorporated by the P2A model.

We train the P2A model on the USC-TIMIT [16] and Haskins Production Rate Comparison [17] EMA datasets. First, we show that the model can generate realistic EMA trajectories, comparable to state-of-the-art (SOTA) A2A models [9, 18]. Second, using the model, we assess the effect of coarticulation by generating EMA trajectories on 9,291 minimal word pairs. Coarticulation effects are most significant for the immediate left and right neighbors of a triggering phoneme (± 1 phonemic distance), and diminish almost threefold at ± 2 distance. Coarticulation effects gradually diminish thereafter but are still significant at ± 7 distance. Coarticulation resistance varies across different places of articulation in consonants. Dentals, alveolars, and postalveolars are more coarticulation resistant than bilabials, labiodentals, and velars. These findings generally align with earlier studies [13], but also suggest a longer-range effect than has been typically found [5, 19, 20], which might relate to the brain's speech planning mechanism [21, 7].

2. Phoneme-to-articulatory Model

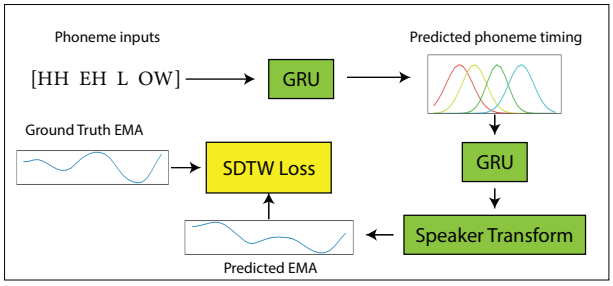

Figure 1: *P2A model architecture*

The P2A model generates an "average" expected EMA time series given any phonemic sequence input. The model has three major components (Figure 1): a bidirectional gated re-

10.21437/Interspeech.2024-2191

current unit (GRU) [22] neural network that predicts the timing of each phoneme, a second bidirectional GRU that generates speaker-independent EMA from the predicted phoneme timing, and a set of speaker-specific linear transforms that generate speaker dependent EMA. The entire model is trained end to end with a soft dynamic time warping (SDTW) [23] loss to account for utterance-specific timing variability that is not possible to predict from a phoneme sequence alone. The code is available at `https://github.com/cffan/ema_coarticulation`.

2.1. Phoneme timing prediction

Given an input phoneme sequence $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, with each $\boldsymbol{x}_i \in \mathbb{R}^c$ representing a one-hot encoding of the ith phoneme, the phoneme timing bidirectional GRU model, denoted as f_θ, predicts the influence of each phoneme at every timestep. We use a Gaussian distribution to model the influence. A similar idea was previously suggested by [24]. Phoneme $\boldsymbol{x}_i$'s Gaussian distribution is parameterized by two values, μ_i and σ_i, which are outputs of the f_θ:

$$\boldsymbol{M} = f_\theta(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}) \quad \boldsymbol{M} \in \mathbb{R}^{n \times 2} \tag{1}$$

where $\mu_i = M_{i,0}$ represents the time point of maximum influence for phoneme i, while $\sigma_i = M_{i,1}$ represents the width of its influence. The Gaussian model permits each time point to be dominated by one phoneme, with nearby Gaussians overlapping to produce coarticulation. $\boldsymbol{M}$ is transformed into $\boldsymbol{M'} \in \mathbb{R}^{t \times n}$ representing the influence of n input phonemes over t timesteps with a Gaussian kernel, i.e. $M'_{k,i} = \exp(-\frac{(k-\mu_i)^2}{2\sigma_i^2})$

2.2. EMA generation

Given the predicted timing $\boldsymbol{M'}$ of the input phonemes, our model proceeds to generate the final EMA time series $\hat{\boldsymbol{Y}}$ as follows:

$$\hat{\boldsymbol{Y}} = h_\lambda(g_\phi(\boldsymbol{M'})) \quad \hat{\boldsymbol{Y}} \in \mathbb{R}^{t \times s} \tag{2}$$

where s is the number of EMA sensor measurements, h_λ denotes a speaker-specific linear transform function, and g_ϕ is a bidirectional GRU that generates speaker-independent EMA representations. Each row $\hat{Y}_{k,:}$ is the generated EMA measurements at time k.

2.3. Model hyperparameters and training

f_θ and g_ϕ are each defined as a 2-layer bidirectional GRU with 128 hidden units. Each speaker has a unique speaker-specific linear transform function h_λ. The smoothness parameter γ in SDTW is set to 1. Hyperparameters are tuned on the validation set. The P2A model is trained end-to-end using Adam optimizer with a learning rate of 4e-4 and batch size of 64 on one NVIDIA A100 GPU. The gradient vector's L2 norm is scaled to 10 to stabilize training.

3. Measuring Coarticulation

To analyze coarticulation, we use minimal word pairs. These are pairs of words that differ in only one phoneme. For instance, "pat [pæt]" and "bat [bæt]" is a minimal pair that differs at the first phoneme. Coarticulation analysis has previously utilized minimal pairs of phonemic sequences [15, 4]. We extend this method to larger sets of words. More generally, given a minimal word pair that differs at position i, we have the following phonemic sequence:

$$[...P_{i-2}\, P_{i-1}\, \boldsymbol{P_i}\, P_{i+1}\, P_{i+2}...]$$

where P_i is the coarticulation trigger, and other phonemes $P_{k|k\neq i}$ are the targets of coarticulation. When comparing the difference between the EMA representations of phonemes at the same position in the pair, we expect the difference to peak at the ith position and gradually drop for further away targets. We use the P2A model to generate EMA trajectories for the pair. Then for each phoneme P_k we time-average the EMA around its maximum influence timestep μ_k to obtain a vector representation $\boldsymbol{\varphi}_k$:

$$\boldsymbol{\varphi}_k = \frac{1}{2\tau+1} \sum_{j=\mu_k-\tau}^{\mu_k+\tau} \hat{Y}_{j,:} \quad \boldsymbol{\varphi}_k \in \mathbb{R}^s \tag{3}$$

where τ is a hyperparameter controlling the width of the phoneme, and s is the number of EMA sensor measurements. We use $\tau = 1$ for all experiments to minimize the influence of nearby phonemes.

We use Euclidean distance to measure the difference between each phoneme's EMA representation in two words *A* and *B* from a minimal pair:

$$d_k = ||\boldsymbol{\varphi}_k^A - \boldsymbol{\varphi}_k^B|| \tag{4}$$

By computing d_k across all phonemic positions for many minimal word pairs, we can obtain a comprehensive and statistically significant view of modeled coarticulation magnitude and temporal extent.

4. Results

4.1. Datasets

We train the model on USC-TIMIT and Haskins Production Rate Comparison (HPRC) datasets. USC-TIMIT is a 4-speaker dataset containing 1 hour of 100 Hz EMA. HPRC is an 8-speaker dataset containing 6.3 hours of 100 Hz EMA. Training on multiple speakers enables the model to learn average speaking patterns. We use a 90%-10% train-validation split for USC-TIMIT, and 80%-20% split for HPRC. Training features include the 2D positions (along the posterior-anterior and inferior-superior axes) and the 2D velocities of each EMA sensor. Adding 2D velocities improves the quality of generated EMA. Features are normalized by speaker-specific mean and standard deviation. Only generated 2D positions are used to evaluate performance and measure coarticulation. Given limited EMA training data, the phoneme input set is simplified by omitting stress markers from vowels, resulting in 39 phonemes and an additional silence phone.

4.2. Quality of generated EMA

Table 1: *Comparing PCC and RMSE on model-predicted EMA (after dynamic time warping) and ground truth EMA on the USC-TIMIT and HPRC datasets. The confidence interval is computed via bootstrap resampling.*

Dataset	PCC [95% CI]	RMSE [95% CI]
USC-TIMIT	0.83 [0.82, 0.84]	0.59 [0.57, 0.62]
HPRC	0.85 [0.85, 0.85]	0.53 [0.53, 0.53]

We compare the model-generated EMA with normalized ground truth EMA data using Pearson correlation coefficients (PCC) and root mean squared error (RMSE). The generated EMA data is aligned with the ground truth using dynamic time warping (DTW) before comparison. Table 1 shows the PCC and RMSE on two datasets. To place these numbers in context, one SOTA A2A model achieves 0.78 PCC on the HPRC dataset [9], and an earlier model achieves 0.61 and 0.50 RMSE on USC-TIMIT and HPRC datasets [18], comparable to ours. This shows that our P2A model can generate high-quality time-warped EMA. We use data from speaker F01 in HPRC dataset for all the results below.

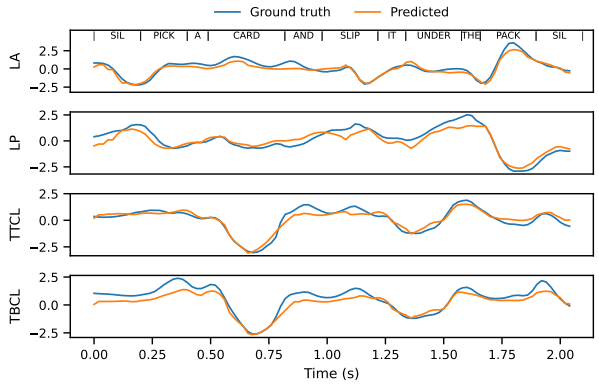

Figure 2: *Comparing tract variables for model-generated vs. ground truth EMA for the sentence "Pick a card and slip it under the pack". Ground truth word alignments are shown in the top panel. The generated EMA is close to the ground truth EMA.*

Figure 2 shows the generated EMA is qualitatively similar to the ground truth on a validation sentence. Four normalized tract variables (TVs) are visualized: lip aperture (LA), lip protrusion (LP), tongue tip constriction location (TCCL), and tongue body constriction location (TBCL). TVs are computed using lip and tongue sensor positions as in [25].

4.3. Coarticulation examples

We use two VCV pairs from [3] to show that our P2A model can generate EMA with known coarticulation patterns. Figure 3 shows the predicted timing (1st row) and four normalized TVs (2nd to 5th row) for each pair. The left panel compares [ʌvi] and [ʌvʌ], and the right compares [ʌzi] and [ʌzʌ]. The tongue position for [i] is more forward than that for [ʌ]. As a result, the TTCL and TBCL for [i] (light red) in both panels have higher values, indicating a more frontal position, than those for [ʌ] (dark red). In both panels, the tongue tip and body positions start to diverge before the onset of the last vowel, showing anticipatory coarticulation. The divergence starts earlier in the left panel than the right because [z] has to be produced with the tongue close to the alveolar ridge, whereas the tongue has less constraint for producing [v]. This result is similar to that in [3], indicating that our P2A model can accurately produce coarticulatory dynamics.

4.4. Temporal extent and magnitude of coarticulation

To comprehensively measure coarticulation, we enumerate 9,291 minimal word pairs from 10,000 most common words (according to Google Books NGram) in the CMU Pronunciation dictionary. For each minimal pair, we use the P2A model to generate their EMA trajectories, extract each phoneme's EMA representation using Eq 3, and measure the EMA representation difference using Eq 4.

Figure 4 shows the temporal extent and magnitude of coarticulation. The y-axis is the magnitude of coarticulation normalized to the average distance between all phonemes (such that a 100% distance means that a phoneme changed so much as to resemble a completely different phoneme). The x-axis is the phonemic distance to the coarticulation trigger. The most significant coarticulation effect (31%±16%, mean±std) occurs on

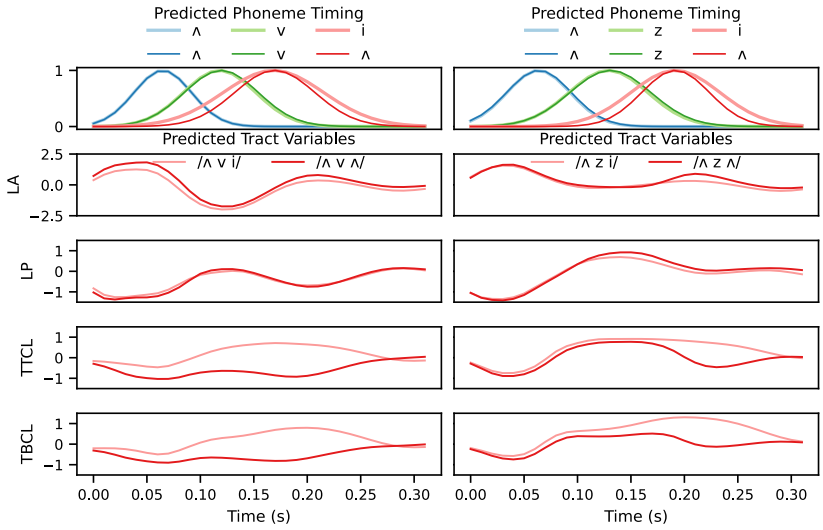

Figure 3: *Model-generated EMA can reproduce known coarticulation on VCV pairs.* ***Left****: predicted phoneme influence and tract variables for [ʌvi] and [ʌvʌ],* ***Right****:[ʌzi] and [ʌzʌ]. TTCL and TBCL diverge before the onset of the last vowel, indicating anticipatory coarticulation. The divergence starts earlier in the left panel as compared to the right, showing the difference in coarticulation resistance between [v] and [z].*

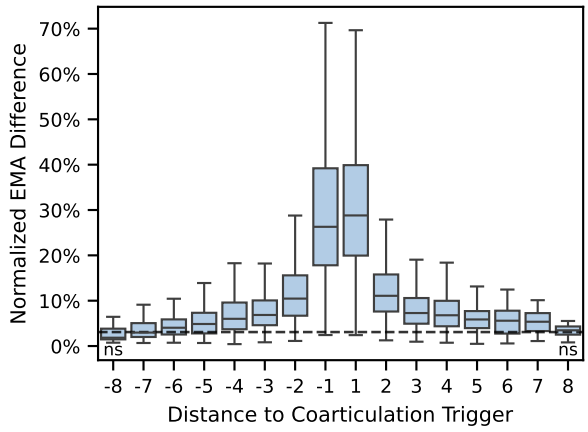

Figure 4: *Quantifying the magnitude and extent of coarticulation. Each box shows the effects of coarticulation at a given phonemic distance to the trigger (whiskers mark the 1.5 interquartile range). The dashed line is the baseline EMA difference. Coarticulation is most pronounced at ±1 distance, but is still significant compared to baseline at ±7 distance.*

the trigger's immediate left and right neighbors (±1). The coarticulation effect reduces significantly to 12%±7% for targets at ±2 distance. For further targets, the coarticulation effect gradually decreases (8% at ±3 distance and 6% at ±4 distance). The wide range of magnitude at each position is due to the varying distances between the trigger phonemes. For instance, trigger pairs such as [ʃ] and [ʧ] cause less coarticulation than [ʃ] and [p].

We used random sentence pairs with either the first or last phonemes changed to measure the baseline magnitude due to model inaccuracy/noise at a far distance, shown as the dashed line. Compared to the baseline, there is still a significant (p=0.01) coarticulation effect at ±7 distance, indicating a longer-range coarticulation effect than has typically been found previously [26, 19, 27]. This result also provides a new justification for the triphone model [28], i.e., the majority of coarticulation can be modeled by considering only the left and right neighbors.

4.5. Coarticulation resistance

Coarticulation resistance refers to the degree to which phonemes resist being affected by adjacent phonemes. Our measure of coarticulation magnitude can be used as a measure of coarticulation resistance (smaller EMA differences indicate a greater resistance to coarticulation). Figure 5 shows the effects of coarticulation on consonants, grouped by place of articulation, at ±1 distance to the trigger. Bilabials, labiodentals, and velars are more affected by coarticulation (lower coarticulation resistance) because producing bilabials and labiodentals does not require the tongue body, and velars can be produced with different tongue constriction locations [29]. In contrast, dentals, alveolars, and postalveolars produced with a raised and fronted tongue dorsum exhibit less coarticulation (higher coarticulation resistance). In one study [3] that analyzed coarticulation resistance of American English consonants, [b], [v], and [g] were found to be less resistant to coarticulation than [ð], [d], [ʒ], and [z]. These results are consistent with ours, suggesting that our framework can be flexibly used to study various aspects of coarticulation.

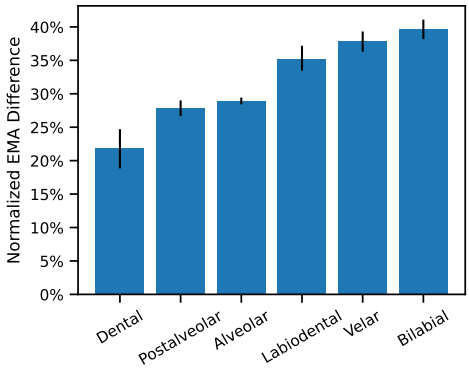

Figure 5: *Comparing the coarticulation resistance of consonants (error bars represent bootstrap-resampled 95% confidence intervals). Dentals, postalveolars, and alveolars are more resistant to coarticulation due to the involvement of the tongue.*

5. Discussion

Ideally, our findings would be validated using real EMA data. However, doing so is difficult, as it would require many repetitions of each word pair (to average over utterance-specific variability) and many word pairs. Searching over the HPRC dataset used in this work, we found only ∼60 word pairs per speaker that had at least 10 repetitions. Additionally, cross-word coarticulation and phoneme alignment errors in real EMA data could confound the results. Alternatively, another future direction for validating this work could be to perform the same analysis on simulated acoustic data using advanced text-to-speech models, which aggregate thousands of hours of audio data [30, 31].

In future work, model-based analysis of coarticulation could provide new insight into coarticulation patterns in different languages, or into the development of coarticulation patterns in children. We intend to use the framework presented here to analyze neural data recorded during continuous speech [32, 33], which could provide insights into the neural mechanism of speech production.

This work is a first step towards a model-based, comprehensive analysis of coarticulation. We leveraged existing EMA datasets and a novel P2A model to generate realistic EMA for many more minimal word pairs than would be realistically feasible for a typical coarticulation experiment. Using the model-generated EMA, we measured the magnitude and temporal extent of coarticulation up to ±8 phonemic distance from the trigger. Our results are consistent with previous studies of coarticulation and suggest a more pronounced long-range coarticulation effect than previously appreciated.

6. Acknowledgements

We thank K. Livescu for her invaluable feedback and insightful comments, and B. Davis, K. Tsou, and S. Kosasih for administrative support. Support was provided by Wu Tsai Neurosciences Institute, Howard Hughes Medical Institute, Larry and Pamela Garlick, Simons Foundation Collaboration on the Global Brain, and NIH-NIDCD R01DC014034.

7. References

[1] D. Recasens, "Coarticulation," in *Oxford research encyclopedia of linguistics*, 2018.

[2] S. E. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[3] C. A. Fowler and L. Brancazio, "Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation," *Language and speech*, vol. 43, no. 1, pp. 1–41, 2000.

[4] Z. Liu, Y. Xu, and F.-f. Hsieh, "Coarticulation as synchronised CV co-onset–parallel evidence from articulation and acoustics," *Journal of Phonetics*, vol. 90, p. 101116, 2022.

[5] M. Grosvald, "Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation," *Journal of Phonetics*, vol. 37, no. 2, pp. 173–188, 2009.

[6] P. S. Beddor, J. D. Harnsberger, and S. Lindemann, "Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates," *Journal of Phonetics*, vol. 30, no. 4, pp. 591–627, 2002.

[7] F. H. Guenther, *Neural control of speech*. Mit Press, 2016.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in matlab," *The Journal of the Acoustical Society of America*, vol. 115, pp. 2430–2430, 2004.

[9] P. Wu, L.-W. Chen, C. J. Cho, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated RNN," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[11] K. Iskarous, C. Mooshammer, P. Hoole, D. Recasens, C. H. Shadle, E. Saltzman, and D. H. Whalen, "The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance," *The Journal of the acoustical society of America*, vol. 134, no. 2, pp. 1271–1282, 2013.

[12] B. Lindblom and H. M. Sussman, "Dissecting coarticulation: How locus equations happen," *Journal of phonetics*, vol. 40, no. 1, pp. 1–19, 2012.

[13] D. Recasens and A. Espinosa, "An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan," *The Journal of the acoustical society of America*, vol. 125, no. 4, pp. 2288–2298, 2009.

[14] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

[15] C. E. Gelfer, F. Bell-Berti, and K. S. Harris, "Determining the extent of coarticulation: Effects of experimental design," *The Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2443–2445, 1989.

[16] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[17] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[18] M. Parrot, J. Millet, and E. Dunbar, "Independent and automatic evaluation of speaker-independent acoustic-to-articulatory reconstruction," in *Proc. Interspeech 2020*, 2020, pp. 3740–3744.

[19] D. Recasens, "Long range coarticulation effects for tongue dorsum contact in VCVCV sequences," *Speech Communication*, vol. 8, no. 4, pp. 293–307, 1989.

[20] H. S. Magen, "The extent of vowel-to-vowel coarticulation in english," *Journal of Phonetics*, vol. 25, no. 2, pp. 187–205, 1997.

[21] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, no. 1, pp. 3–35, 1990.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[23] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *International conference on machine learning*. PMLR, 2017, pp. 894–903.

[24] A. Löfqvist, "Speech as audible gestures," in *Speech production and speech modelling*. Springer, 1990, pp. 289–322.

[25] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion," in *Proc. Interspeech 2019*, 2019, pp. 859–863.

[26] S. Ahmed and M. Grosvald, "Long-distance vowel-to-vowel coarticulation in Arabic: Influences of intervening consonant pharyngealization and length," *Language and Speech*, vol. 62, no. 2, pp. 399–424, 2019.

[27] P. West, "Long-distance coarticulatory effects of british English /l/ and /r/: An EMA, EPG and acoustic study," in *Proccedings of the 5th Seminar on Speech Production: Model and Data*, 2000, pp. 105–108.

[28] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. USA: Prentice Hall PTR, 2000.

[29] J. Dembowski, M. J. Lindstrom, J. R. Westbury, M. Cannito, K. Yorkston, and D. Beukelman, "Articulator point variability in the production of stop consonants," *Neuromotor speech disorders: Nature, assessment, and management*, pp. 27–46, 1998.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. Interspeech 2017*, 2017, pp. 4006–4010.

[31] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[32] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.

[33] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.