ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096800

ARTICULATION GAN: UNSUPERVISED MODELING OF ARTICULATORY LEARNING

Gašper Beguš[1*], *Alan Zhou*[2*], *Peter Wu*[1†], *Gopala K. Anumanchipalli*[1†]

[1]University of California, Berkeley, [2]Johns Hopkins University

ABSTRACT

Generative deep neural networks are widely used for speech synthesis, but most existing models directly generate waveforms or spectral outputs. Humans, however, produce speech by controlling articulators, which results in the production of speech sounds through physical properties of sound propagation. We introduce the Articulatory Generator to the Generative Adversarial Network paradigm, a new unsupervised generative model of speech production/synthesis. The Articulatory Generator more closely mimics human speech production by learning to generate articulatory representations (electromagnetic articulography or EMA) in a fully unsupervised manner. A separate pre-trained physical model (ema2wav) then transforms the generated EMA representations to speech waveforms, which get sent to the Discriminator for evaluation. Articulatory analysis suggests that the network learns to control articulators in a similar manner to humans during speech production. Acoustic analysis of the outputs suggests that the network learns to generate words that are both present and absent in the training distribution. We additionally discuss implications of articulatory representations for cognitive models of human language and speech technology in general.

Index Terms— articulatory phonetics, unsupervised learning, electromagnetic articulography, deep generative learning

1. INTRODUCTION

Humans produce spoken language with articulatory gestures [1]. Sounds of speech are generated by airflow from the lungs passing through articulators, which causes air pressure fluctuations that constitute sounds of speech. The main mechanism in speech production is thus control of the articulators and airflow [1]. During language acquisition, children need to learn to control articulators and produce articulatory gestures such that the generated sounds correspond to the sounds of language they are exposed to.

This learning is complicated by the fact that sound is an entirely different modality compared to articulatory gestures. Children need to learn to control and move articulators from sound input without direct access to the articulatory data of their caregivers. While some articulators are visible (such as lips and tongue tip, jaw movement), many are not (vocal folds, tongue dorsum). There is debate on whether spoken language acquisition is fully unsupervised due to direct and indirect negative evidence [2]. Articulatory learning, however, is likely fully unsupervised. Caregivers ordinarily do not provide any explicit feedback about articulatory gestures to language-acquiring children.

Most models of human speech production output audio data of speech without articulatory representations. In actual speech, however, humans control articulators and airflow, while a separate physical process results in sounds of speech.

To build a more realistic model of human spoken language, we propose a new deep learning architecture within the GAN framework [3, 4, 5, 6, 7, 8]. In our proposal, the decoder (synthesizer or the Generator network) learns to output approximates of human articulatory gestures while never accessing articulatory data. The generated articulatory gestures are represented with thirteen channels that match the twelve channels used to record human articulators during electromagnetic articulography (EMA) plus an additional channel for voicing. The generated articulatory movements are then passed through a separate *physical model* of sound generation that takes articulatory channels and converts them into waveforms. This physical model is taken from a pre-trained EMA-to-speech model (ema2wav) which transforms electromagnetic articulography into speech waveforms [9]. This physical model component is a model of physical sound propagation and is cognitively irrelevant, which is why its weights are not updated during training.

Articulatory learning in this model needs to happen in a fully unsupervised manner. The Articulatory Generator needs to transform random noise in the latent space into the thirteen channels such that the independent pre-trained EMA-to-speech physical model will generate speech. The Discriminator receives waveform data synthesized based on the Articulatory Generator's generated channels. The Generator in our model never directly accesses articulatory data. Like humans, it needs to learn to control articulators without ever directly accessing them (e.g. vocal folds or tongue dorsum are never visible during speech acquisition). The only information available to humans during acquisition and our model during training is the auditory feedback from the perception component of speech that corresponds to the Discriminator network in our model.

1.1. Prior work

Speech synthesis from articulatory representations has recently been performed using deep neural networks [10, 11, 12, 13, 14, 9]. The objective in most existing proposals, however, is to synthesize waveforms from articulatory representations in a supervised setting, rather than a fully unsupervised generation of the articulatory representations themselves. [15, 16] proposes an autoencoder model that learns to encode and decode between motor parameters and auditory representations in an unsupervised manner. However, this model trains both the encoding and decoding aspects of the model simultaneously, and focuses on the relationship between auditory representations and a motor latent space. By contrast, our GAN model is trained with a static pretrained articulatory model similar to how children learn to speak with a full set of articulators. In addition, rather than decoding back and forth between motor and auditory information, our model is able to generate articulatory parameters directly by sampling from a general-purpose latent space. To our knowledge, this paper presents the first architecture in

*Gašper Beguš and Alan Zhou contributed equally to this work. *Corresponding author: Gašper Beguš (begus@berkeley.edu).*

†G.K.A. and P.W. are supported by NSF #2106928.

which a generative model learns to produce unprompted articulatory gestures that result in speech in a fully unsupervised way.

Computational models of language almost always disregard the articulatory component. Currently, articulatory phonology is a proposal that comes closest to modeling linguistic representations from articulatory representations [17, 18], and phonological structure can be inferred from articulatory data [19]. However, these models take articulatory gestures as a given (as measured on human subjects) and do not model unsupervised learning and generation of articulatory gestures from auditory feedback.

A model of unsupervised articulatory learning is not only a more realistic representation of human speech, but is useful for conducting cognitive simulations that have the potential to reveal which properties of speech emerge because of articulatory factors and which properties are cognitively conditioned [20]. In engineering application, learning to generate plausible articulatory gestures with accompanied synthesized speech is useful for lip synchronization [21] (with potential applications in robotics or gaming industry). The modelling of articulatory information has also been identified as being useful in the detection of audio deepfakes [22]. Generation of articulatory gestures is thus potentially useful for creating more realistic speech synthesis technologies, as well as providing another adversarial approach that deepfake detectors can use to improve their accuracy.

2. THE MODEL

Our articulatory model takes the architecture of WaveGAN [5], and replaces its Generator with a combination of an Articulatory Generator and a physical model of articulation. The Articulatory Generator is a modification of the WaveGAN Generator that maps random noise to 13 channels of time-series data corresponding to articulatory representations and voicing. The physical model is a pretrained autoregressive encoder that maps the modified Generator's articulatory output into speech data. Note that the weights of the physical model are frozen during training: we constrain the problem so that the Articulatory Generator learns to produce articulatory movements that will result in realistic speech.

2.1. Articulatory Generator

The Articulatory Generator G is adapted from the Generator network from WaveGAN [5]. It takes as input a latent noise vector z and uses 5 layers one-dimensional transpose convolutions to upsample the noise into waveform data $G(z)$. Unlike WaveGAN, our Articulatory Generator generates 13 channels (one channel of voicing plus the x- and y-axis for 6 articulators). Due to the physical model's low sample rate, the dimensionality of each layer is also lower than in WaveGAN, with individual dimensionalities of $32 \times 512, 64 \times 512, 128 \times 256, 128 \times 256, 256 \times 13$, respectively.

2.2. Physical Model

We take the EMA-to-speech encoder trained on MNGU0 from [9] to be a physical model of articulation $\mathcal{A}$. This autoregressive model takes as input 13 channels of time-series and outputs a 16 kHz waveform corresponding to speech. Specifically, the 13 channels of articulatory features include one channel of voicing, plus the x and y coordinate positions each of the lower incisor, upper lip, lower lip, tongue tip, tongue body, and tongue dorsum.

3. TRAINING

We train our model using the same WGAN-GP scheme [23] as in [5], except we replace the Generator's output with the output of our two-step articulatory inference:

$$\max_D \min_G V(D,G) = \mathbb{E}_{x\sim P_x}[D(x)] - \mathbb{E}_{z\sim P_z}[D(\mathcal{A}(G(z)))]$$

where P_x and P_z are the training and noise distributions, and the Discriminator D is constrained to be 1-Lipshitz function.

We train the network on 8 words from TIMIT [24]: *ask*, *dark*, *year*, *water*, *wash*, *rag*, *oily*, and *greasy* for 354,200 training steps with a batch size of 8. These specific words were chosen because of their relatively equally frequent appearance in TIMIT.[1] We limit the number of training words to facilitate learning as well as to mimic language acquisition more closely: productive vocabulary size is relatively small at the initial stages of language acquisition [25].

The training data for the Generator is different from the MNGU0 data set [26] training dataset used in the EMA-to-speech physical model (which involves a single speaker of British English). This mimics human language acquisition, where children need to learn from multiple adults while having a single set of articulators. Our training is additionally complicated by MNGU0 and TIMIT involving speakers of different varieties of English (British vs American).

We additionally train an unmodified WaveGAN network [5] on the same 8 words from TIMIT [24] as a baseline to compare against our articulatory model. This model was trained for 138,600 steps with a batch size of 32.

4. RESULTS

4.1. Performance

To test how well the ArticulationGAN performs compared to WaveGAN, we randomly generated 200 outputs from the ArticulationGAN and WaveGAN models (400 total). A trained phonetician who is not a coauthor was hired to annotate and transcribe the outputs in order to avoid potential bias and to account for noise in the outputs. The outputs were annotated as (i) intelligible words of English, (ii) intelligible sequences of sounds that are not words of English, and (iii) unintelligible outputs. The results are given in Figure 1. Intelligible outputs include all sounds that were transcribable by the trained phonetician; the proportion of comfortably intelligible outputs is likely lower. The WaveGAN model performs slightly better on the intelligibility task (87% vs. 72%), but the ArticulationGAN outputs a higher proportion of intelligible outputs (words and nonwords) that are not part of training data (innovative outputs).

The weaker intelligibility of ArticulationGAN is likely due to its more difficult training objective. The WaveGAN Generator only needs to produce outputs that are themselves similar to the training distribution drawn from TIMIT. On the other hand, the Articulatory Generator must produce outputs that approximate the TIMIT data after being passed through a fixed articulatory model that has been trained on MNGU0 data. As previously mentioned, TIMIT data is drawn from 630 speakers across eight dialects of American English [24], while MNGU0 data is drawn from a single speaker of British English [26].

[1] Counts of each token as well as generated EMA and waveform data, annotations, and checkpoints are available at `doi.org/10.17605/OSF.IO/X37HA`. The code is available at `github.com/gbegus/articulationGAN`.

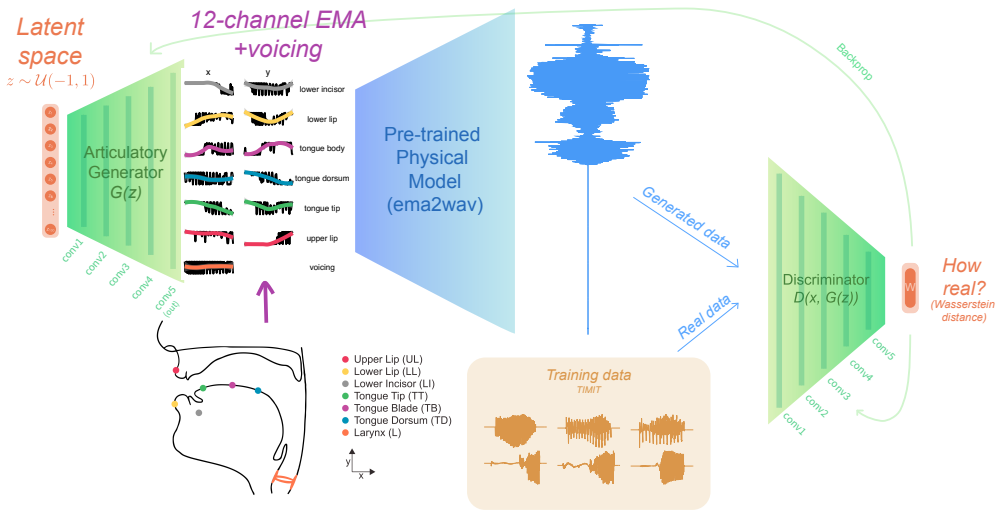

Fig. 1. The architecture of the ArticulationGAN. The Articulatory Generator takes 100 latent variables z and generates 12 EMA channels and the channel for voicing. The pre-trained physical model (ema2wav) takes the generated EMA and transforms them into waveforms.

Model	Intelligible	Unintelligible	Innovative
WaveGAN	174 (87%)	26 (13%)	87 (50%)
ArticulationGAN	143 (72%)	57 (29%)	110 (77%)

Table 1. Counts of annotated outputs in WaveGAN and ArticulationGAN architectures. The 200 annotated words per model are divided into intelligible and unintelligible outputs. The Innovative column indicates those intelligible outputs (words and non-words) that are not part of training data. 33 (17%) outputs are training data words in ArticulationGAN (compared to 87 or 44% in WaveGAN).

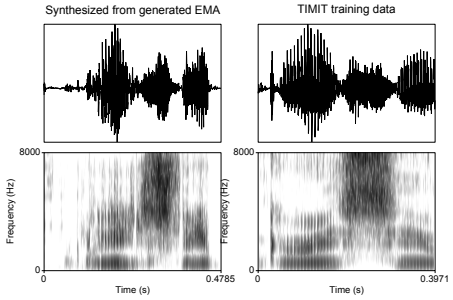

Fig. 2. Generated output *greasy* and its corresponding (TIMIT) datapoint used during training.

Nevertheless, ArticulationGAN produces a higher amount of innovative data compared to both the TIMIT and MNGU0 datasets. The results suggest that the ArticulationGAN not only learns words that are represented in both TIMIT training data and MNGU0 dataset (e.g. *wash*), but also words that are absent from the MNGU0 dataset and the TIMIT training dataset. For example, the ArticulationGAN generated outputs that were transcribed as *wash* ['wɔʃ], *fast* ['fæst], *greasy* ['gɹisi], and *coffee* ['kɔfi]. *Wash* is part of TIMIT and MNGU0 training data. *Fast* and *coffee* are only present in the MNGU0 data. *Fast* is acoustically close to *ask* in the training data. *Coffee* is distant to its closest equivalent in the TIMIT training data (*greasy*), but *greasy* is absent from MNGU0. The *ema2wav* model is never trained to generate *greasy* from EMA, yet our ArticulationGAN generates several outputs that can be reliably transcribed as *greasy* (Figure 2).

We also observe overrepresentation of *w*-initial words in the 200 outputs of the ArticulationGAN compared to TIMIT training data (OR $= 1.53, p < 0.01$), but not in WaveGAN outputs (OR $= 0.94, p = 0.74$). Gestures for [w-] are easier to acquire compared to other initial consonants. It appears that ease of articulation of labial consonants plays a role in articulatory learning in our models (similar to language acquisition [27, 28]).

4.2. Analyzing generated gestures

To analyze unsupervised learning of articulatory gestures in the ArticulationGAN model, we compare real (MNGU0) and generated (ArticulationGAN) EMA channels and corresponding acoustic outputs (waveforms). We analyze articulatory gestures in two generated outputs transcribed as *wash* ['wɔʃ] and *fast* ['fæst]. These words were chosen because *wash* is present in both TIMIT training and MGNU0 data, while *fast* is an innovative output. Because *greasy* is fully absent from MNGU0 training data, we cannot compare generated and real EMA for this word.

Figure 3 illustrates the 12 channels plus voicing for *wash*. We observe the network learns relatively stable articulatory targets, except during transitions between targets or when an articulator does not play an active role for a given phoneme sequence. For example, the x axis of tongue dorsum and the lower incisor position do not play a central role in the articulation of *wash*, which is why this channel is relatively noisy in Figure 3.

To interpret articulatory gestures and compare real human EMA to generated EMA, we visualize x and y-axis values in 2D space for each electrode placement. Because the Generator has no restrictions that would penalize rapid changes (as is the case in human muscle and movements), we smooth the generated EMA with LOESS smoothing. Figure 4 contains generated and real EMA for *wash* and an innovative output *fast*. Tongue tip and lower/upper lip are the most relevant articulators for *wash* and *fast*. We observe very similar

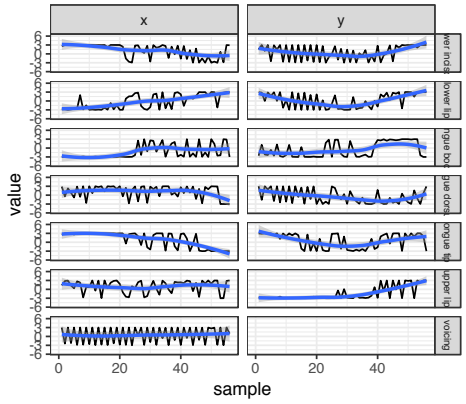

Fig. 3. Generated EMA channels and voicing for *wash* with LOESS smoothing.

gestures between GAN and EMA for *wash*, and an almost identical pattern the lower lip gestures for *fast*.

4.3. Quantitative comparison between generated and real EMA

We further performed a quantative comparison between gestures for the generated and real EMA. To account for differences in timing, we perform dynamic time warping (DTW) between smoothed generated EMA and real EMA for each dimension (x and y) and each electrode placement. We then compute Pearson's product-moment correlation (r) on two time series data for each channel to estimate the time-aligned correlation between generated and real EMA.

	wash		fast	
Place	x	y	x	y
tongue tip	0.70	0.90	0.99	0.96
tongue body	0.94	0.91	0.32	0.79
lower lip	-0.52	0.70	0.85	0.94
upper lip	0.51	0.90	0.64	0.43
lower incisor	0.87	0.66	0.31	0.72
tongue dorsum	0.41	0.91	0.24	0.89

Table 2. Pearson's product-moment correlation (r) for *wash* and *fast* after DTW alignment of two time series.

The quantitative comparison in Table 2 reveals a high degree of correlation in gestures between real EMA and GAN-generated EMA. Tongue tip gestures in *fast* are almost identical ($r = 0.99$ in x-axis and $r = 0.96$ for y axis). We observe that tongue tip, lower lip, and tongue body have highest correlations and the y-axis is better correlated than the x-axis. This is expected as vertical movements are more consequential in these words.

4.4. Limitations & future directions

Despite the training complexities discussed in Section 1, the intelligibility of ArticulationGAN's outputs is not substantially lower than that of WaveGAN (Table 1). ArticulationGAN outputs a higher proportion of innovative intelligible outputs. This is not unexpected from cognitive modeling perspective: speech production (articulatory learning) is substantially more difficult than speech perception (acoustic learning), and innovative outputs are common during articulatory learning.

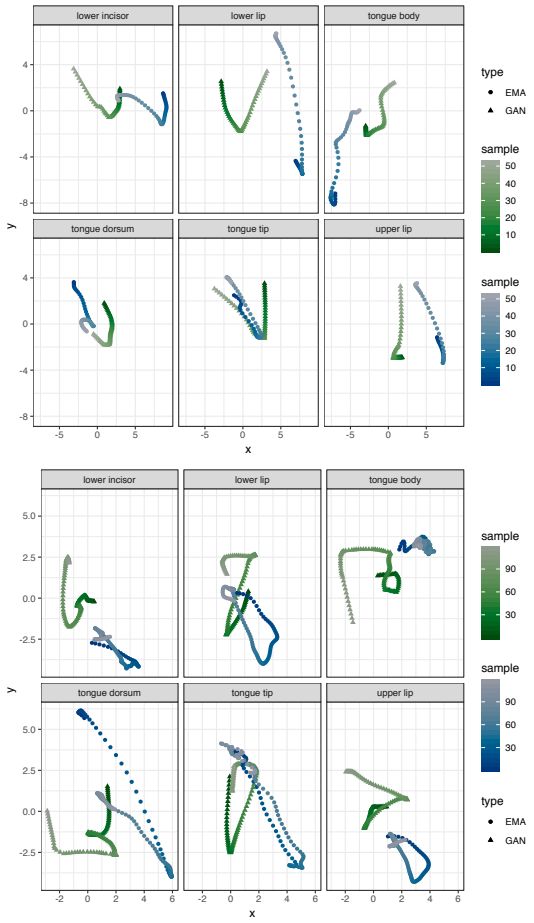

Fig. 4. Real EMA channels (blue circles) and smoothed, generated EMA (green triangles) in 2D space for output transcribed as *wash* (top) and *fast* (bottom). Real EMA is multiplied by 3.0 for comparison. Temporal dimension (sample) is represented with shading. Note the similarity in trajectories between generated and real channels especially for tongue tip, lower lip, and tongue body. Quantitative analysis of trajectories in Table 2 shows a high degree of correlation

EMA data is a very low-dimensional representation of articulation in human speech. Adding articulatory representations (e.g. more channels or additional articulation data types) might improve performance and provide higher resolution insights about articulation. Also, our model operates with a single Discriminator and a single 5-layer Generator that needs to generate 13 1D channels, which may impact performance. Adding multiple subdiscriminators has also been shown to increase performance in the GAN framework [29].

5. CONCLUSION

This paper proposes a new model for unsupervised learning of articulatory gestures in human speech production. To our knowledge, we present the first fully unsupervised deep generative network that learns to generate articulatory representation from latent noise based exclusively on the audio inputs. We argue that the Articulatory Generator learns to generate human-like articulatory representations and propose a technique to quantitatively estimate the similarities.

6. REFERENCES

[1] Kenneth N. Stevens, *Acoustic phonetics*, Current studies in linguistics 30. MIT Press, Cambridge, MA, 1998.

[2] Barbara C. Lust, *Child Language: Acquisition and Growth*, Cambridge University Press, 2006.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. 2014.

[4] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ArXiv 1511.06434*, 2015.

[5] Chris Donahue, Julian J. McAuley, and Miller S. Puckette, "Adversarial audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[6] Gašper Beguš, "Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks," *Frontiers in Artificial Intelligence*, vol. 3, pp. 44, 2020.

[7] Gašper Beguš, "CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks," *Neural Networks*, vol. 139, pp. 305–325, 2021.

[8] Gašper Beguš and Alan Zhou, "Modeling speech recognition and synthesis simultaneously: Encoding and decoding lexical and sublexical semantic information into speech with no direct access to speech data," in *Proc. Interspeech 2022*, 2022, pp. 5298–5302.

[9] Peter Wu, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala Krishna Anumanchipalli, "Deep Speech Synthesis from Articulatory Representations," in *Proc. Interspeech 2022*, 2022, pp. 779–783.

[10] Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert, "Robust articulatory speech synthesis using deep neural networks for BCI applications," in *Proc. Interspeech 2014*, 2014, pp. 2288–2292.

[11] Sandesh Aryal and Ricardo Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[12] Yu-Wen Chen, Kuo-Hsuan Hung, Shang-Yi Chuang, Jonathan Sherman, Wen-Chin Huang, Xugang Lu, and Yu Tsao, "Ema2s: An end-to-end multimodal articulatory-to-speech system," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[13] Marc-Antoine Georges, Jean-Luc Schwartz, and Thomas Hueber, "Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE," in *Proc. Interspeech 2022*, 2022, pp. 774–778.

[14] Marc-Antoine Georges, Pierre Badin, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber, "Towards an articulatory-driven neural vocoder for speech synthesis," in *ISSP 2020 - 12th International Seminar on Speech Production*, Providence (virtual), United States, Dec. 2020.

[15] Yashish M. Siriwardena, Guilhem Marion, and Shihab Shamma, "The Mirrornet: Learning audio synthesizer controls inspired by sensorimotor interaction," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 946–950.

[16] S. Shamma, P. Patel, S. Mukherjee, G. Marion, B. Khalighinejad, C. Han, J. Herrero, S. Bickel, A. Mehta, and N. Mesgarani, "Learning Speech Production and Perception through Sensorimotor Interactions," *Cerebral Cortex Communications*, vol. 2, no. 1, 2020.

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

[18] Caitlin Smith, Charlie O'Hara, Eric Rosen, and Paul Smolensky, "Emergent gestural scores in a recurrent neural network model of vowel harmony," in *Proceedings of the Society for Computation in Linguistics 2021*, Feb. 2021, pp. 61–70.

[19] Pallavi Baljekar, Sunayana Sitaram, Prasanna Kumar Muthukumar, and Alan W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *Proc. Interspeech 2015*, 2015, pp. 3194–3198.

[20] Gašper Beguš, "Distinguishing cognitive from historical influences in phonology," *Language*, vol. 98, no. 1, pp. 1–34, 2022.

[21] Xiaohong Li, Xiang Wang, Kai Wang, and Shiguo Lian, "A novel speech-driven lip-sync model with CNN and LSTM," in *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2021, pp. 1–6.

[22] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor, "Who are you (I really wanna know)? detecting audio DeepFakes through vocal tract reconstruction," in *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, Aug. 2022, pp. 2691–2708, USENIX Association.

[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems 30*, pp. 5767–5777. 2017.

[24] J. S. Garofolo, Lori Lamel, W. M. Fisher, Jonathan Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1993.

[25] Larry Fenson, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis, and Joan Stiles, "Variability in early communicative development," *Monographs of the Society for Research in Child Development*, vol. 59, no. 5, pp. i–185, 1994.

[26] Korin Richmond, Phil Hoole, and Simon King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," *Interspeech 2011*, p. 1505–1508, 2011.

[27] Bénédicte de Boysson-Bardies and Marilyn May Vihman, "Adaptation to language: Evidence from babbling and first words in four languages," *Language*, vol. 67, no. 2, pp. 297–319, 1991.

[28] Sharynne McLeod and Kathryn Crowe, "Children's consonant acquisition in 27 languages: A cross-linguistic review," *American Journal of Speech-Language Pathology*, vol. 27, no. 4, pp. 1546–1571, 2018.

[29] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 17022–17033.