Deciphering Assamese Vowel Harmony with Featural InfoWaveGAN

Sneha Ray Barman[1], *Shakuntala Mahanta*[1,2], *Neeraj Kumar Sharma*[1,3]

[1]Centre for Linguistic Science & Technology, [2]Department of Humanities & Social Sciences,
[3]Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology Guwahati, Guwahati-781039, India

{sneha.barman, smahanta, neerajs}@iitg.ac.in

Abstract

Traditional approaches for understanding phonological learning have predominantly relied on curated text data. Although insightful, such approaches limit the knowledge captured in textual representations of the spoken language. To overcome this limitation, we investigate the potential of the Featural InfoWaveGAN model to learn iterative long-distance vowel harmony using raw speech data. We focus on Assamese, a language known for its phonologically regressive and word-bound vowel harmony. We demonstrate that the model is adept at grasping the intricacies of Assamese phonotactics, particularly iterative long-distance harmony with regressive directionality. It also produced non-iterative illicit forms resembling speech errors during human language acquisition. Our statistical analysis reveals a preference for a specific [+high,+ATR] vowel as a trigger across novel items, indicative of feature learning. More data and control could improve model proficiency, contrasting the universality of learning.

Index Terms: Computational Linguistics, Generative Adversarial Networks, Phonological Learning, Vowel Harmony, Assamese Language

1. Introduction

Understanding human language acquisition involves examining various aspects of grammar, with phonology playing a crucial role. Approaches such as Optimality Theory [1] models phonological grammar as input-output pairs, selecting the most optimal output based on input. In contrast, Maximum Entropy Grammar [2] proposes that grammar constraints are language-specific and learned from data, employing machine learning techniques to determine these constraints.

Integrating neural network methodologies, multiple studies have recently proposed Generative Adversarial Network (GAN) and WaveGAN models for phonetic and phonological learning, with advancements like Featural InfoWaveGAN (fiwGAN) enhancing lexical learning and demonstrating proficiency in phonological representation learning across languages. A comprehensive review of prior generative linguistics studies is provided by Pater [3]. Beguš [4] proposed Generative Adversarial Network (GAN [5]) as a method to model learnability more organically, wherein phonetic and phonological processes are represented as mappings from randomly distributed latent space to generative data. Donahue et al. [6] introduced the WaveGAN model for audio data, employing Deep Convolutional GAN (DCGAN [7]) architecture to learn language features from continuous speech signals. This model processes 1-second audio files sampled at 16 kHz, converting them into vectors for input to the discriminator network. WaveGAN, adapted for phonetic and phonological learning [4], effectively captures linguistic representations through a one-to-one mapping from latent space to generated data, mimicking how humans construct underlying phonological representations from speech streams. However, human acquisition also involves lexical learning, the ability to store unique information associated with meaning-bearing units or phonemes while adhering to language-specific phonotactics. Beguš [8] enhanced the WaveGAN architecture by integrating a lexical learning component, the Q-network [9], resulting in Featural InfoWaveGAN (fiwGAN). This advancement introduces a new latent space structure to comprehend featural representations of phonetic and phonological learning. The fiwGAN generator network learns to produce acoustic data by encoding unique lexical information and generating innovative acoustic data associated with distinct binary codes for each lexical item. Examining the fiwGAN network's performance on English allophonic distribution, Beguš & Zhou [10] demonstrated its generative and learning capabilities by producing data comparable to human speech acquisition outputs. Furthermore, Chen & Elsner [11] evaluated FiwGAN's robustness by assessing English and French nasality, confirming its ability to learn phonological representations where latent variables correspond to identifiable phonological features. This highlights FiwGAN's proficiency as a phonological learner, which is particularly evident in controlled settings when comparing the same feature across languages.

Building on these studies, we investigate the potential of the Featural InfoWaveGAN model to learn iterative long-distance vowel harmony using raw speech data. Vowel harmony is a phonological phenomenon observed in many languages where vowels within a word tend to be influenced by one another in terms of their phonetic properties, such as tongue height, tongue position, or roundedness [12, 13]. Owing to this phenomenon, certain vowels in a word may undergo changes to better match the phonetic characteristics of neighboring vowels. The result is an enhancement of the overall harmony and fluidity of pronunciation within the word [14]. Vowel harmony can be categorized as regressive, where vowels are influenced by following vowels or progressive, where vowels are influenced by preceding vowels. We focus on Assamese, a language known for its phonologically regressive and word-bound vowel harmony [15, 16]. We demonstrate that the model is adept at grasping the intricacies of Assamese phonotactics, particularly iterative long-distance harmony with regressive directionality.

2. Vowel Harmony in Assamese

Assamese is an Indo-Aryan language spoken in the northeastern state of Assam (in India)by 15.3 million individuals [17]. The variety analyzed in this paper is the colloquial Assamese spoken in the eastern region of Assam, also known as Upper Assam.

This features eight surface vowels, namely, [i, e, ɛ, ɑ, ɔ, o, ʊ, u], and twenty consonants, namely [p, pʰ, b, bʰ, t, tʰ, d, dʰ, k, kʰ, g, gʰ, m, n, ŋ s, z, x, h, ɹ, j, w, l] [15].

Table 1: *Vowel inventory of Assamese*

Vowels	Front	Back	ATR
High	i	u	+ATR
		ʊ	-ATR
Mid	e	o	+ATR
	ɛ	ɔ	-ATR
Low		ɑ	-ATR

In Assamese, the surface vowel occurrences are restricted in the following ways: *(a)* The two high vowels /i/ and /u/ are pronounced with an advanced tongue root [+ATR], as are the mid vowels /e/ and /o/; *(b)* The mid vowels / ɛ/ and / ɔ/ are slightly lower than /e/ and /o/ and are not realized with an advanced tongue root [-ATR], and *(c)* The vowels /e/ and /o/ do not occur word-finally or initially, they occur only under circumstances of vowel harmony.

Table 2: *Examples of vowel harmony in Assamese [15]*

Stem	Gloss	-suffix	Harmonised	Gloss
/pɛt/	'belly'	-u	[petu]	'pot-bellied'
/pɑgɔl/	'mad-M'	-i	[pɑgoli]	'mad-F'
/gʊl/	'to mix'	-i	[guli]	'I mix'
/bɔsɔr/	'year'	-i	[bosori]	'yearly'

Assamese exhibits right-to-left ATR harmony where the high vowels /i/ and /u/ trigger [+ATR] harmony of [-ATR] high /ʊ/ and mid vowels [/ɛ/, /ɔ/] in the language. From the examples of Table 2, we can argue that the stem vowels alternate under the influence of suffixes in Assamese. [+ATR] acts as a dominant feature that spreads from the suffix vowels in the leftward direction and affects vowels on the left side of the triggering vowel. As seen in the examples, harmony applies to all the [-ATR] vowels preceding a [+ATR] vowel within the word domain showcasing long-distance iterative harmony (see example 4 from Table 2).[15] proposed that the harmony sequence in Assamese may be subject to the tendency of avoiding featural cooccurrence of [-ATR, +ATR] vowels. Therefore, [-ATR] vowels do not trigger harmony playing a recessive role. /ɑ/ appears to be an opaque vowel that blocks harmony from its targeting vowels to its left (zʊnɑki 'firefly' *zunɑki). Exceptional patterns emerge when /ɑ/ in the root morpheme is raised, followed by derivational suffixes [uwɑ] and [ijɑ]. /ɑ/ is raised according to the preceding [-ATR] vowel; for example, when the preceding vowel is /ɛ/, /ɑ/ becomes [e] (ɛlah-elehuwɑ) and [o] otherwise. [16] proposed that iterative harmony is gradient and vowels closer to the trigger is affected more than the ones further away from the trigger. Our study attempts to navigate these claims through modeling learnability computationally.

3. Proposed Approach

The novelty of our study lies in computing iterative long-distance harmony from raw speech in an unsupervised manner. Vowel harmony is a complex phonological process requiring the learner to grasp crucial aspects like directionality, domains, features, iteration, locality, and opacity [18]. The vowel harmony system of Assamese provides a crucial case study for directionality as it is neither stem-controlled nor dominant-recessive

[15]. We train the model with Assamese speech data containing both harmonic and non-harmonic words. We quantify the presence of [+ATR] harmony when the frequency of F1 of the target vowel [-ATR] is much lower in the vicinity of the trigger [+ATR] vowel than respective F2 and F3 [19, 20]. Regressive directionality is anticipated when V2 explains the properties of V1 in a V1CV2 setting. We analyze the output audios statistically and observe that although the Generator strings vowels and consonants together to form a novel word, it does follow a vowel sequence within the word domain. We also identify that some innovative items have lexical meaning in Assamese, suggesting the emergence of lexical learning from the fiwGAN model.

4. Model Architecture

The fiwGAN neural network model comprises three primary components: a Generator, a Q-network, and a Discriminator (refer to Figure 1). Unlike traditional GAN models, where the Generator typically takes uniformly distributed latent variables as input, fiwGAN's Generator operates within a latent space composed of binary codes (ϕ) and latent variables (z). Here, each binary variable (ϕ_n) corresponds to a distinct feature [21]. Trained to maximize the Discriminator's error rate while minimizing the Q-network's error, the Generator produces a vector containing approximately 16384 data points for a 1 second of audio data (sampled at 16 kHz). These generated outputs, along with real acoustic data, are then assessed by the Discriminator, which computes the Wasserstein distance [22] between the generated and real data.

In parallel, the Q-network, which shares the same input data as the discriminator, operates differently. Its final layer comprises nodes corresponding to categorical variables (ϕ). Unlike the Discriminator, the Q-network's loss function drives updates not only to its own weights but also to those of the Generator. This mechanism enables the Generator to associate each lexical item with a unique latent code, allowing the Q-network to retrieve the lexical information solely from the acoustic signals. This process facilitates lexical learning within the model. While the Generator produces raw acoustic data similar to but not an exact replication of the real input data, it effectively captures essential characteristics of the acoustic signals. [1]

5. Materials

5.1. Assamese Speech Data

For this experiment, 15 native Assamese speakers (7 males and 8 females) between the ages of $19 - 35$ were consulted. The speech data was recorded in the soundproof booth at the Phonetics and Phonology laboratory of the Indian Institute of Technology Guwahati (India) using a Tascam DR-100 MKII recorder. The created speech dataset consisted of 82 spoken words(40 harmonic and 42 non-harmonic)[2]. Each target word was in a carrier sentence written in Assamese script, that is, *"moi X buli kolu"*, which corresponds to "I say X" in English. The participants were asked to utter each sentence at least four times. The collected speech data was then curated through manual segmentation using the audio-visual utility in PRAAT[23] software.

---

[1] We used the Pytorch implementation of fiwGAN made available by [21] at `https://github.com/gbegus/ciwganFiwGAN-pytorch.git`.

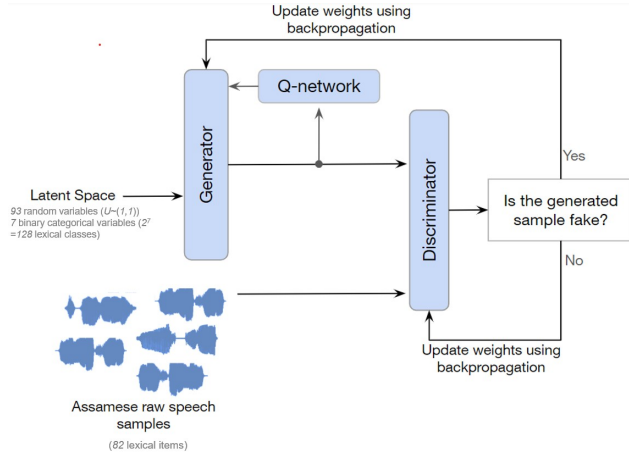[2] The dataset is available at `https://github.com/sneha2599/IS2024_fiwGAN-for-vowel-harmony.git`

Figure 1: *An illustration of fiwGAN architecture used in this work. We chose N = 82 spoken utterances corresponding to lexical items in the Assamese language. The latent space contains 93 uniformly distributed latent variables, z, and 7 binary features (ϕ) for 2^7=128 lexical classes.*

This resulted in 4789 clean speech utterances (or tokens), each corresponding to an Assamese word.

Table 3: *Example dataset for training fiwGAN*

Stem	-suffix	Surface	Harmony type
dilɛ	-i	dilei	harmonic
nokorilɛ	-u	nokorileu	harmonic
gɔrɔm	-ɔ-t	gɔrɔmɔt	non-harmonic
bɛpɑr	-i	bɛpɑri	non-harmonic

5.2. Model Implementation

The fiwGAN model takes 4789 (3169 harmonic and 1620 non-harmonic) 1-second-long unannotated single-channel audio files sampled at 16 kHz in raw waveforms. The latent space consists of 93 uniformly distributed latent variables z ($U \sim (-1,1)$) and 7 binary variables, allowing for 2^7=128 unique classes, represented as [1,0,0,0,0,0,0 for word1], [0,1,0,0,0,0,0 for word2], etc. The Generator and Discriminator were trained with Adam optimizer while the Q-network was trained with RM-SProp algorithm at a .0001 learning rate with a batch size of 64. It ran for 72 hours on an NVIDIA RTX A4500 GPU at the CLST lab of IIT Guwahati. The Generator data was analyzed following 960 training epochs ($\sim$ 44000 steps). Based on their intelligibility and noise quality, 64 out of 100 outputs were examined in our experiment after manual listening and segmentation in Praat[23]. We extracted the first formants of the vowels in training and generated data at 10-time points using Praat and analyzed only the mean value to quantify the presence of ATR vowel harmony[23], followed by regression analysis in R [24] to assess the learnability of directionality and other discrete categories related to vowel harmony.

6. Results

On the completion of each training epoch, the Generator produces 100 audio files. In previous studies exploring En-

glish allophonic distributions and French nasality [8, 11], the Generator learned to produce speech-like sounds after close to 650 epochs. In our case, the speech-like sounds were found to be generated after 720 epochs. Intelligible sounds were generated after 800 epochs. After 960 epochs, the model started generating lexical items identical to the training dataset ([prohori],[ɛkɛbɑrɛ],[pɔlɔx]) and innovative outputs ([dekhisi],[korobe],[korisuwɑ],[debeku] etc.). Additionally, some illicit forms ([nɔkorilu], [pɔdobi], etc.) were also found in the outputs. An illustration of some outputs is shown in Figure 2.

Subsequently, we analyzed the innovative outputs and observed that the model strings together vowels and consonants from the training data; for example, [dekhisi] is assumed to be influenced by the word [dekhisu] in the input, while [korobe] may stem from [korobi]. In the context of vowel harmony, the novel items [debeku], [dekhisi], and [korisuwɑ] follow the vowel harmony pattern where [+high,+ATR] vowels [i,u] trigger harmony of the underlying [-ATR] vowels turning them into [+ATR]. We observed that [e]/[o] occurs when followed by [i]/[u] locally [korisuwɑ/dekhisi] as well as over a longer domain [debeku]. This suggests the capability of the model to learn the phonotactics of Assamese and the iterative long-distance vowel harmony.

Analyzing the illicit outputs, we observed that the vowels in them follow a specific pattern where the trigger vowels impact the immediately preceding vowel, suggesting non-iterative harmony([pɔdobi] instead of [podobi]) (see Figure 3). Moreover, [korisuwɑ] is close to a real-world compound word [kori suwɑ'do see'] and a word the model has not encountered in the input, implying the emergence of lexical learning from the training.

6.1. Statistical Analysis

We examined the generated outputs statistically to quantify the presence of discrete categories related to vowel harmony in Assamese The formant values of vowels in the training data were analyzed. It was observed that the first formant frequency (F1) of the [-ATR] vowels is much lower in the vicinity of the [+ATR] vowels (see Figure 3). We hypothesize that if V2 in the V1CV2 setting explains V1 better than V1 explains V2, the dataset follows regressive directionality. If not, the directionality is assumed to be left-to-right. A linear mixed-effects regression analysis was carried out on the training dataset [25] (using R [24]). This helped examining the relationship between the dependent variable F1V1 and predictor variables V1 and V2 [3]. V1 and V2 are the fixed effects, while 'word' is the random effect in the model. The likelihood ratio test compares the full model to a null model. The AIC score revealed a statistically significant difference between the two models (AIC(full)=1588.8, AIC(null)=1595.9). The chi-square test indicates a significant difference in model fit between the null and full models (χ^2=33.062, df=13, p<0.001), reflecting a significant effect of V2 on the F1 value of V1 (see Table 4). The dataset is then categorized into two subsets: harmonic and non-harmonic. Both the subsets are analyzed using the lmer model. The fixed-effects results from the full model for harmony [4] showed a distinction between V1, V2, and the intercept representing F1 of V1. The estimate of the intercept is about 552.129 Hz with a p-value<2e-

---

[3] V1full.model=lmer(F1V1$\sim$V1+V2+(1|word), data=formants, REML=FALSE)

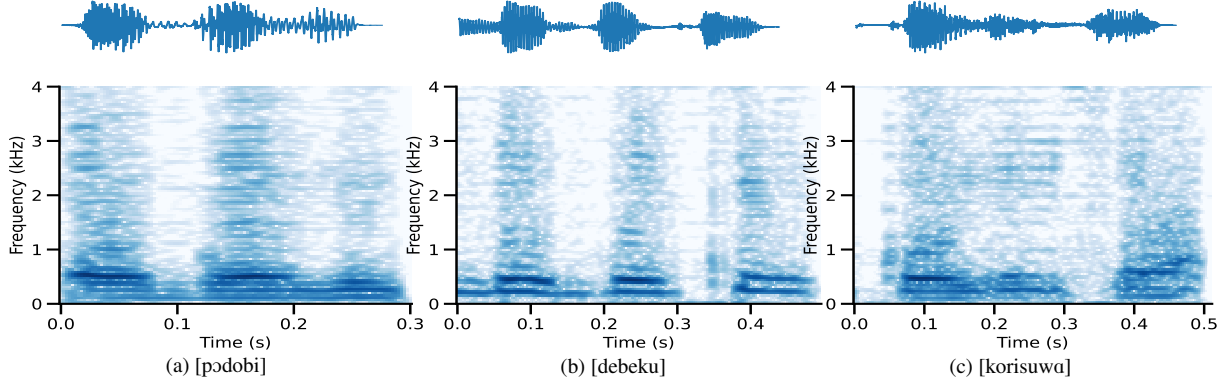[4] V1vhfull=lmer(F1V1$\sim$V1+V2+(1|word),data=har, REML=FALSE)

Figure 2: *Spectrograms of three fiwGAN generated audio files: (a)[pɔdobi], also an illicit item; (b) [debeku], also an innovative item following long-distance iterative harmony; and (c)[korisuwɑ] a novel word with lexical meaning in the Assamese language.*
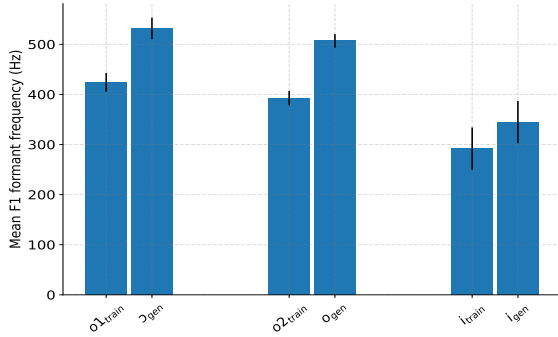

Figure 3: *F1 comparison of [podobi] (training data; shown in bars) and [pɔdobi] (generated data; shown in hatched bars). Here, o1 and o2 denote the first and second vowel, and i denotes the third vowel, in the input training data "podobi".*

16. A comparison of the full and null model using the likelihood ratio test showed that V2 affects V1 (χ^2= 27.829, df = 7, p = 0.0002361), which reveals that the contrast is statistically significant (AIC(full)=467.04, AIC(null)=480.87)(see Table 4). Assuming the results from statistical models for training data as a baseline, the generated items are annotated and then fit to linear regression models in R to examine the relationship between the dependent variable F1V1 and predictor variables V1 and F1V2 (see Table 5). The model was statistically significant (F(6, 15) = 9.504, p = 0.0002087), indicating that at least one predictor variable had a non-zero effect. The Adjusted R-squared value of 0.7084 suggests that the model explained approximately 70.84% of the variability in F1V1. The coefficients for individual predictor variables were significant at various levels. Our regression model suggests that V1 and, to some extent, F1V2 are associated with variations in F1V1. Further analysis shows that the coefficients of V2[T.i] are significantly higher than other V2 variables (Estimate=-279.11, t-value=-3.376,p= 0.00817). This supports the idea that [i], a [+high,+ATR] vowel in Assamese, acts as a triggering vowel in the machine-generated outputs, suggesting the learnability of height as a trigger feature for Assamese vowel harmony. The overall analysis suggests that the model learns the right-to-left harmony pattern, which is evident in the influence of V2 on V1

for both first and second formant frequencies.

Table 4: *LMER model for the training dataset*

Data	Directionality	Fixed effects	DF	χ^2	p
Whole	right-to-left	F1V1~V1+**V2**	13	33.062	<0.001
	left-to-right	F1V2~V2+**V1**	10	6.5156	0.77
Only [+ATR]	right-to-left	F1V1~V1+**V2**	7	27.829	< 0.001
	left-to-right	F1V2~V2+**V1**	2	1.6522	0.43

Table 5: *Linear regression model for machine-generated items*

Data	Estimate	t-value	p-value
Whole	605.25	7.793	< .001
only V2[i] coefficient	-279.11	3.376	.01

7. Conclusion

In summary, this study presented long-distance vowel pattern learning, emphasizing learners' acquisition of intricate vowel changes crucial for grasping vowel harmony. Unlike previous literature focusing solely on phonological learning, our findings mark an advancement, leveraging deep learning models to facilitate unsupervised learning of both acoustic and articulatory properties alongside underlying segmental features and vowel harmony. Through analysis of the Generator performance across epochs and manipulated latent space variables, insights into human-like speech generation have emerged. Our observations also indicate that, given limited data, the fiwGAN is an effective phonotactic learner, which is evident from its ability to learn complex systems like iterative long-distance harmony. Analysis of innovative items revealed instances of long-distance harmony alongside error items displaying non-iterative local harmony. The dominance of V2[i] as a trigger vowel indicates the model's capacity to learn trigger features, while statistical analysis suggests grammatical outputs following regressive directionality [5]. Moreover, lexical learning emerged post-training. Notably, no results with the opaque vowel [ɑ] were observed, prompting further inquiry into the model's ability to identify vowel opacity.

---

[5][26] observed coarticulation propensity in Assamese non-harmonic sequences which is strictly phonetic and not a goal of this paper.

8. References

[1] A. Prince and P. Smolensky, "Optimality Theory: Constraint interaction in generative grammar," *Optimality Theory in phonology: A reader*, pp. 1–71, 2004, publisher: Wiley Online Library.

[2] B. Hayes and C. Wilson, "A maximum entropy model of phonotactics and phonotactic learning," *Linguistic inquiry*, vol. 39, no. 3, pp. 379–440, 2008, publisher: MIT press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . .

[3] J. Pater, "Generative linguistics and neural networks at 60: Foundation, friction, and fusion," *Language*, vol. 95, no. 1, pp. e41–e74, 2019.

[4] G. Beguš, "Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks," *Frontiers in artificial intelligence*, vol. 3, p. 44, 2020, publisher: Frontiers Media SA.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[6] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[8] G. Beguš, "Identity-based patterns in deep convolutional networks: Generative adversarial phonology and reduplication," *Trans.Association for Computational Linguistics*, vol. 9, pp. 1180–1196, 2021.

[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[10] G. Beguš and A. Zhou, "Interpreting intermediate convolutional layers of generative CNNs trained on waveforms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3214–3229, 2022.

[11] J. Chen and M. Elsner, "Exploring How Generative Adversarial Networks Learn Phonological Representations," *arXiv preprint arXiv:2305.12501*, 2023.

[12] J. Blevins, *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press, 2004.

[13] S. Rose and R. Walker, "Harmony systems," *The handbook of phonological theory*, pp. 240–290, 2011.

[14] J. J. Ohala, "Towards a universal, phonetically-based, theory of vowel harmony," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 1994, pp. 491–494.

[15] S. Mahanta, *Directionality and locality in vowel harmony: With special reference to vowel harmony in Assamese*. Netherlands Graduate School of Linguistics, 2008.

[16] D. B. Archangeli and J. Yip, "Assamese vowels and vowel harmony," *Journal of South Asian Languages and Linguistics*, vol. 6, no. 2, pp. 151–183, 2020, publisher: De Gruyter.

[17] Government General of India, "Census of India," 2011.

[18] D. Archangeli and D. Pulleyblank, "Harmony," in *The Cambridge Handbook of Phonology*, ser. Cambridge Handbooks in Language and Linguistics, P. d. Lacy, Ed. Cambridge University Press, 2007, pp. 353–378.

[19] T. H. Mintz, R. L. Walker, A. Welday, and C. Kidd, "Infants' sensitivity to vowel harmony and its role in segmenting speech," *Cognition*, vol. 171, pp. 95–107, 2018.

[20] P. Olejarczuk, M. A. Otero, and M. M. Baese-Berk, "Acoustic correlates of anticipatory and progressive [ATR] harmony processes in Ethiopian Komo," *Journal of Phonetics*, vol. 74, pp. 18–41, 2019.

[21] G. 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, publisher: Elsevier.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proc. Intl. Conf. Machine Learning (ICML)*, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 214–223.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[24] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org/

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[26] I. Dutta, I. S., P. Gogoi, and P. Sarmah, "Nature of Contrast and Coarticulation: Evidence from Mizo Tones and Assamese Vowel Harmony," in *Proc. Interspeech 2017*, 2017, pp. 224–228.