Unsupervised modeling of vowel harmony using GAN

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

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

Introduction

- Vowel harmony requires the learner to grasp crucial aspects like **directionality**, **domains**, **features**, **iterativity**, **locality**, and **opacity**.
- Prosodic factors like morphological units such as root and stem significantly influence vocalic alternation (Krämer, 1999).
- Assamese is one of the few Indian languages that exhibit **phonologically regressive** and **word-bound** vowel harmony.
- Our paper investigates **the learnability of vowel harmony in Assamese by a Generative Adversarial Network (GAN) architecture without any external prosodic cues related to it**.

Literature review

- Traditional approaches for understanding phonological learning have predominantly relied on curated text data and supervised training.
- Considering human language acquisition is best modeled unsupervised from raw speech, we take raw unannotated acoustic waveforms as the principal input for training the model.
- WaveGAN (Donahue et al., 2018) learns phonetic and phonological representations similar to how humans construct underlying phonological representations by listening to a speech stream in a language.
- FiwGAN (Beguš, 2020), a fusion of WaveGAN and InfoGAN, proposes a new latent space to simultaneously learn **featural representations** of phonetic and phonological learning.
- The Generator network learns how to generate acoustic data that **encodes unique lexical information and outputs innovative acoustic data as a one-to-one mapping**.

Vowel harmony in Assamese

- Assamese exhibits right-to-left regressive ATR harmony where the high vowels /i/ and /u/ trigger [+ATR] harmony in [-ATR] high and mid vowels in the language, i.e., /ɛ/, /ɔ/ and /ʊ/, except the [−ATR] low vowel /ɑ/ (see Table 1).
- An otherwise opaque vowel [ɑ] becomes either /e/ or /o/ when followed by [-ijɑ] or [-uwɑ].

Assamese	Gloss	-suffix	Harmonised	Gloss
/pɛt/	'belly'	-u	[petu]	'pot-bellied'
/ʊpɔr/	'above'	-i	[upori]	'in addition'
/kɔr/	'do'	-i	[kori]	'I do'
/pɑgɔl/	'mad-M'	-i	[pɑgoli]	'mad-M'

Table 1. Illustrated examples of vowel harmony in Assamese

Training dataset

Root	-suffix	Surface	Harmony type
gɔrɔm	-ɔ-t	gɔrɔmɔt	non-harmonic
dilɛ	-i	dilei	harmonic
nokorilɛ	-u	nokorileu	harmonic
zʊnɑk	-i	zʊnɑki	non-harmonic

Table 2. 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 dataset comprised 82 spoken words(40 harmonic and 42 non-harmonic). Each target word was in a carrier sentence written in Assamese script, that is, *"moi X buli kolu"*, corresponding to, "I say X" in English. The participants were asked to utter each word at least four times. The collected speech data was then curated through manual segmentation using the audio-visual utility in Praat software. This resulted in 4789 clean speech utterances (or tokens), each corresponding to an Assamese word.

Architecture of fiwGAN trained on Assamese

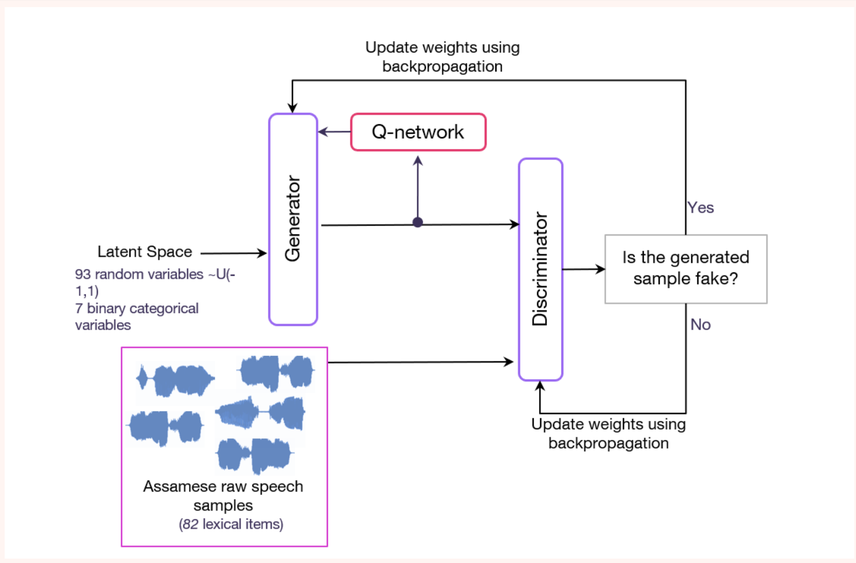


Figure 1. An illustration of fiwGAN architecture used in this work. We chose $N = 82$ spoken utterances corresponding to lexical items in the Assamese language. The latent space contains 93 uniformly distributed latent variables, z, and 7 binary features (ϕ)for 2^7=128 lexical classes.

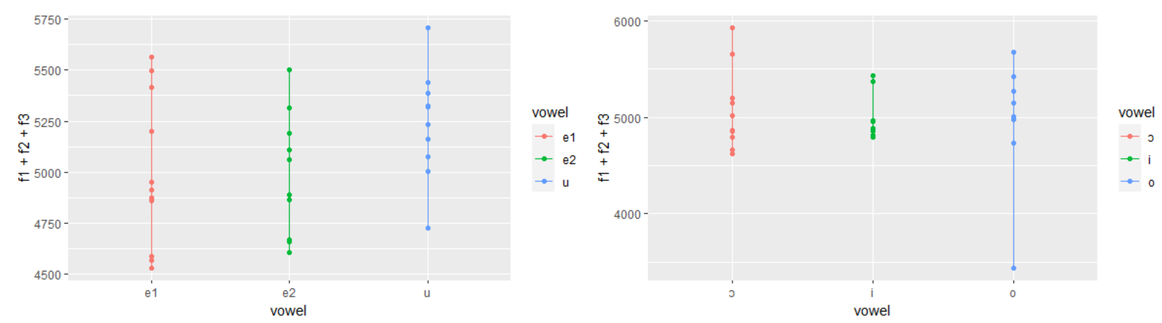
Method

- The Generator was trained on unannotated audio data lasting 1s or less, with a sampling rate of 16 kHz.
- The Generator and Discriminator were trained with Adam optimizer while the Q-network was trained with RMSProp algorithm at a .0001 learning rate with a batch size 64.
- The model produces human-like intelligible speech after 700 epochs; the Generator data is loaded after 920 epochs.
- 100 audio files are generated at each epoch. We analyzed 64 out of 100 outputs generated at 920 epochs.
- The outputs are manually segmented in Praat.

Results

The model generates items that are-

- identical to the training dataset ([**prohori**],[ɛkɛbɑrɛ],[**pɔlɔx**]) [54%] of the dataset],
- innovative ([**dekhisi**],[**korobe**],[**korisuwɑ**],[**debeku**] etc.) (see Figure 2(a)) [45%] of the dataset]
- also, illicit ([**nɔkorilu**], [**pɔdobi**] etc.) (see Figure 2(b))



((a)) [debeku] (innovative item) ((b)) [pɔdobi](illicit item)

Figure 2. First three formants of vowels in two of the generated items

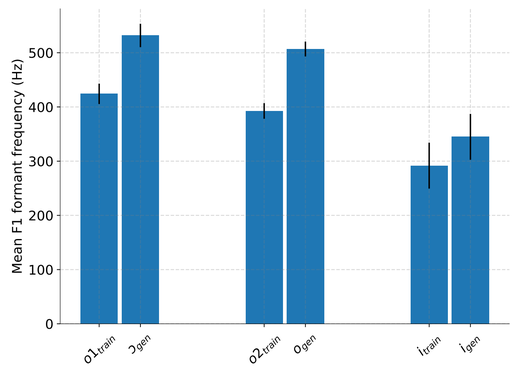


Figure 3. F1 comparison of [podobi] (training data; shown in bars) and [pɔdobi] (generated data; shown in hatched bars). Here, o1 and o2 denote the first and second vowel and i denotes the third vowel, in the input training data *"podobi"*.

Statistical Analysis

- Statistical analysis was carried out to quantify directionality, harmony type, and features of target and trigger.
- Right-to-left directionality was followed in harmonic subsets of both input and output datasets.
- The F1 value of underlying [-ATR] vowels is lowered in the vicinity of [+high, +ATR] vowels (see Figure 1), reflecting vowel harmony (see Table 3).
- The model learns to identify [i] as a trigger vowel with a significantly high coefficient value (see Table 3).

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 3. Results from LMER model for the training dataset

Data	Estimate	t-value	p-value
Whole	605.25	7.793	<.001***
only V2[i] coefficient	-279.11	3.376	<.05**

Table 4. Results from linear regression model for machine-generated items

Contribution, Limitations & Future Scope

Contribution & limitation

- mental representation of linguistic objects (latent space variables) is emerging from raw linguistic input contrary to Chmoskian "Universal Grammar".
- Learning a complex phenomenon like vowel harmony from actual speech data rather than written transcriptions is a big leap forward.
- Further exploration will enhance GAN's ability to model human learning more closely.
- We did not observe any examples of opacity learning in the results.
- The model took longer epochs to generate human-speech-like sounds for Assamese than earlier experiments in English.

Future scope

- **Future questions:** Can the model learn trans-word utterances, thereby identifying word boundaries? Why does it not produce words with the opaque vowel? Which underlying features are responsible for the learning?
- **Future work:** Visualizing Generator loss curve across epochs to understand the pace of learning, generating more outputs in lesser epochs, classifying features responsible for learning vowel harmony in both training and generated dataset, interpreting intermediate layers of the neural networks.

References

Beguš, G. (2020). Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in artificial intelligence*, 3:44.

Donahue, C., McAuley, J., and Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.

Krämer, M. (1999). On the role of prosodic and morphological categories in vowel harmony and disharmony: a correspondence approach. In *Proceedings of ConSOLE*, volume 7, pages 183–199.