

# Modeling unsupervised learning of regressive vowel harmony with GAN: A view from Assamese

Sneha Ray Barman<sup>1</sup>   Shakuntala Mahanta<sup>1,2</sup>   Neeraj Kumar Sharma<sup>1,3</sup>

<sup>1</sup>Centre for Linguistic Science & Technology, <sup>2</sup>Department of Humanities & Social Sciences, <sup>3</sup>Mehta Family School of Data Science & Artificial Intelligence  
Indian Institute of Technology Guwahati, Assam, India  
sneha.barman@iitg.ac.in

## Introduction

- Vowel harmony requires the learner to grasp crucial aspects like **directionality**, **domains**, **features**, **iterativity**, **locality**, and **opacity**.
- Assamese is one of the few Indian languages that exhibit **phonologically regressive** and **word-bound** vowel harmony.
- As a very first attempt to model vowel harmony unsupervised from raw speech, our paper explores the **Featural InfoWaveGAN** model.

## Background

- Existing approaches to modeling phonological learnability have predominantly relied on curated text data and supervised training.
- We take an unsupervised approach with raw, unannotated acoustic waveforms as the principal input for modeling learnability.
- WaveGAN learns phonetic and phonological representations similar to how humans construct underlying phonological representations by listening to a speech stream in a language.
- FiwGAN (Beguš, 2021), 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 to **encode unique lexical information and outputs innovative acoustic data as a one-to-one mapping between randomly distributed latent space and generative data**.

## 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 [/ɨ/] and mid vowels [/ɛ/, /ɔ/], except the [-ATR] low vowel /a/ (see Table 1) Mahanta (2008).
- An otherwise opaque vowel [a] becomes either /e/ or /o/ when followed by [-ija] or [-uwa].

Assamese	Gloss	-suffix	Harmonised	Gloss
/pet/	‘belly’	-u	[petu]	‘pot-bellied’
/upɔr/	‘above’	-i	[upori]	‘in addition’
/kɔr/	‘do’	-i	[kori]	‘I do’
/pagəl/	‘mad-M’	-i	[pagoli]	‘mad-M’
/misa/	‘lie’	-l-ija	[misolija]	‘liar’

Table 1. Illustrated examples of vowel harmony in Assamese

## Training dataset

Root	-suffix	Surface	Harmony type
gɔrɔm	-ɔ-t	gɔrɔmɔt	non-harmonic
dile	-i	dilei	harmonic
nokorile	-u	nokorileu	harmonic
zɔnak	-i	zɔnaki	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

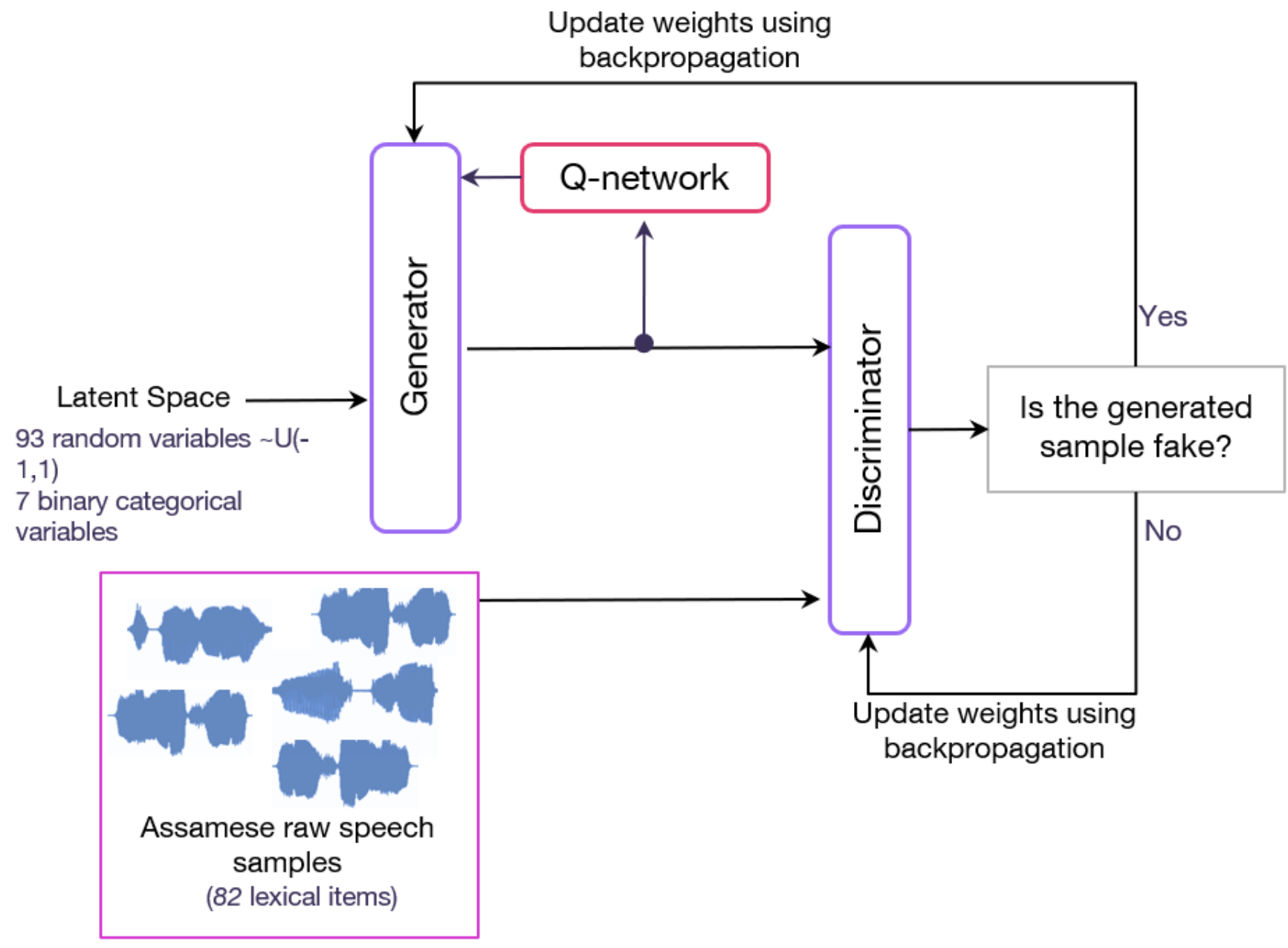


Figure 1. An illustration of fiwGAN architecture used for this work. We chose  $N = 82$  spoken utterances corresponding to lexical items in the Assamese language. The latent space contains 93 uniformly distributed latent variables ( $z$ ) and 7 binary features ( $\phi$ ) for  $2^7=128$  lexical classes.

## 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**],[**ɛkɛbarɛ**],[**pɔlɔx**]) [54%] of the dataset,
- innovative ([**dekhisi**],[**korobe**],[**korisuwa**],[**debeku**] etc.) (see Figure 2(a)) [45%] of the dataset]
- also, illicit ([**nɔkorilu**], [**pɔdobi**] etc.) (see Figure 2(b))

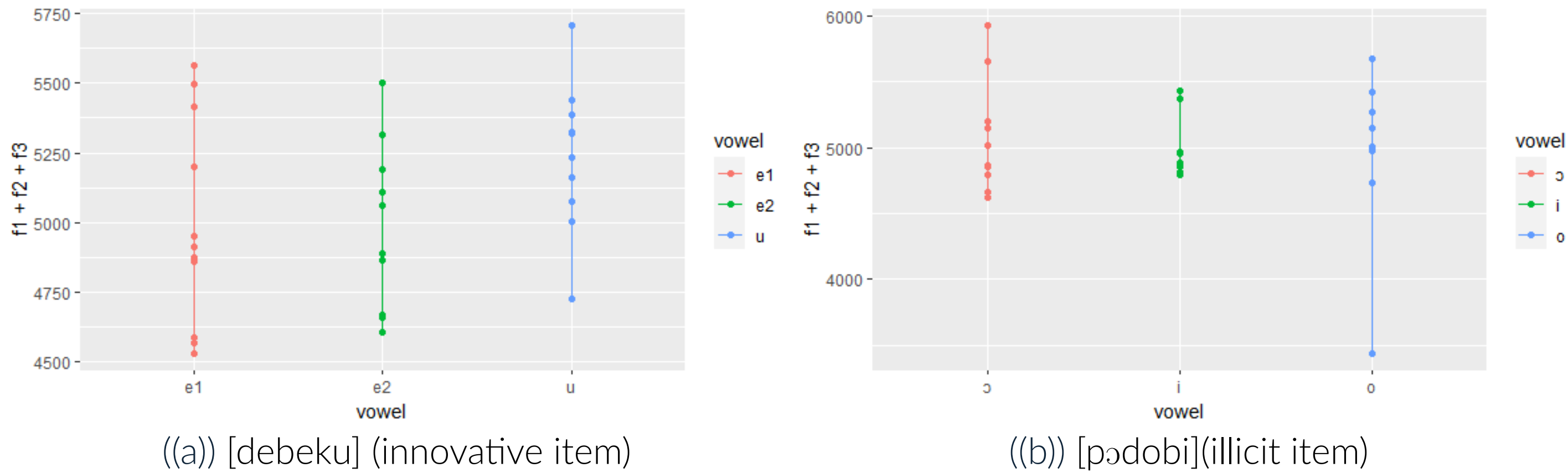


Figure 2. First three formants of vowels in two of the generated items

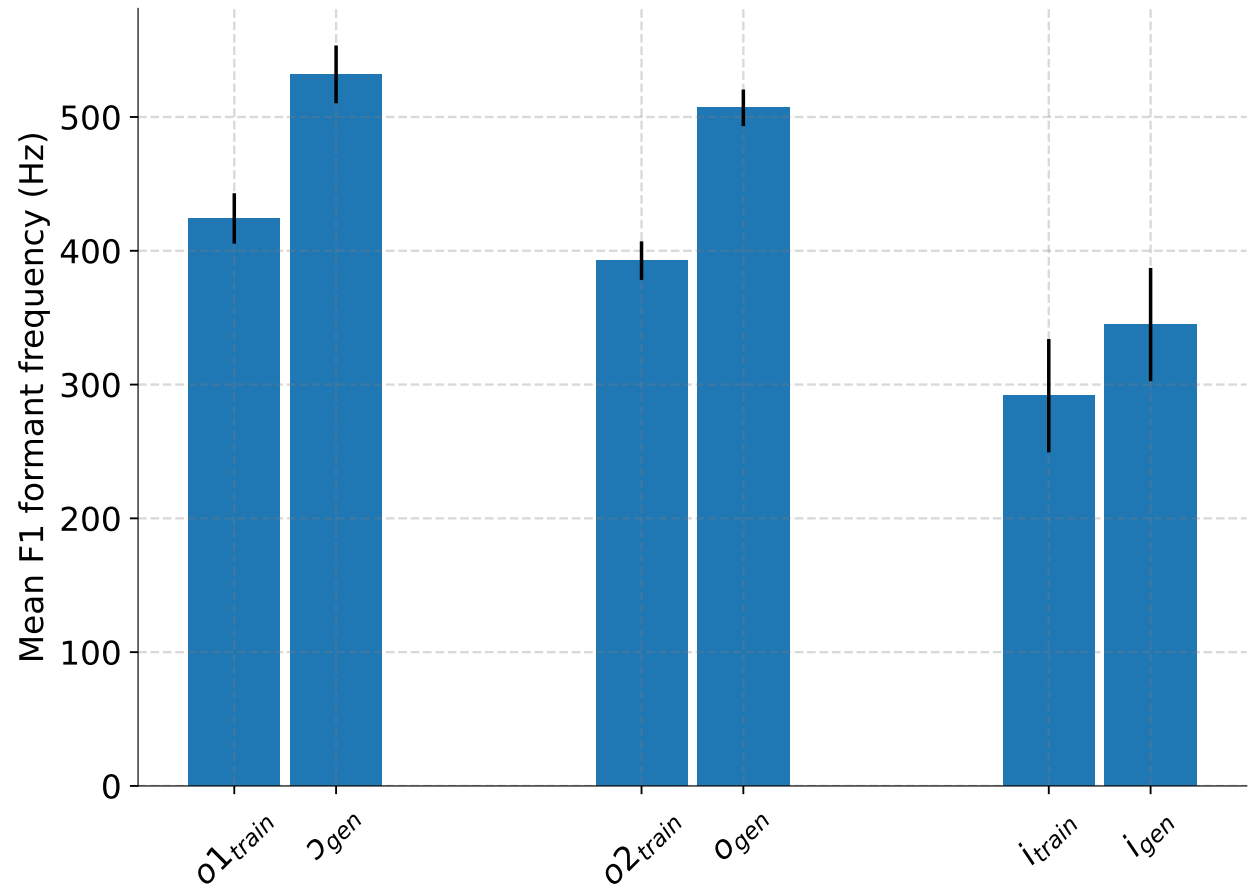


Figure 3. F1 comparison of [podobi] (training data; shown in bars) and [pɔdobi] (generated data; shown in hatched bars). Here, o1 and o2 denote the first and second vowel and i denotes the third vowel, in the input training data ‘*podobi*’.

## 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 ATR 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	$\chi^2$	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) emerges from raw linguistic input.
- Computation of long-distance vowel harmony. Lesser instances than local harmony imply difficulty in learning.
- The illicit items resemble speech errors during human language acquisition.
- Lack of illicit items with opaque vowel /a/.
- Further exploration will enhance GAN's ability to model human learning more closely.
- The model took longer epochs to generate human-speech-like sounds for Assamese than earlier experiments in English and French.

### 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.

## Selected References

- Beguš, G. (2021). Ciwgan and fiwgan: Encoding information in acoustic data to model lexical learning with generative adversarial networks. *Neural Networks*, 139:305–325.
- Mahanta, S. (2008). *Directionality and locality in vowel harmony: With special reference to vowel harmony in Assamese*. Netherlands Graduate School of Linguistics.  
Dataset and results available at: <https://github.com/sneha2599/FiwGAN-Assamese.git>