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

Sneha Ray Barman[1], Shakuntala Mahanta[1], and Neeraj Kumar Sharma[1]
[1]*Centre for Linguistic Science and Technology, Indian Institute of Technology Guwahati*

Introduction: Models of phonological learnability aim to understand how children acquire the phonological grammar of their native language using computational algorithms. To explore how speech interacts with grammatical construct, we examine whether a Generative Adversarial Network (GAN [1]) can capture regressive vowel harmony patterns when trained unsupervised on raw acoustic data. In previous learnability models exploring vowel harmony (Maximum Entropy grammar [2], simple RNN [3]), the learning is supervised, or at least one of either phonetic or phonological learning is assumed to have occurred already. We train the featural InfoWaveGAN model (fiwGAN [4]) with Assamese speech data. Assamese is one of the few Indian languages that exhibits rich vowel harmony. [+high, +ATR] vowels [i, u] trigger right-to-left harmony of [-ATR] vowels [ɛ, ɔ, ʊ] resulting in [e], [o], and [u], respectively [5,6] (see Table 1). FiwGAN comprises three deep convolutional networks: a Generator, a Q-network, and a Discriminator. The Generator, a five-layer convolutional network, is trained to increase the Discriminator's error and Q-network's success rates (see Fig. 1). It is trained to associate lexical items so the Q-network can retrieve lexical code from acoustic signals only, resulting in lexical learning. We analyze the generated items and examine the model's capacity to grasp underlying features like directionality, locality, iteration, and opacity essential to learning vowel harmony. We probe that the model strings elements from the training data to generate an output without external cues. This is similar to [7] 's observation of infants taking vowel harmony as a cue to word segmentation. Our study of modeling learnability computationally also stands on par with the universal nature of human learning.

Materials and model implementation: We recorded 15 native Assamese speakers repeating 82 target words [CVCV(C)(V)] at least four times within the phrase "[moi X buli kolu]" ("I say X"), yielding 5000 tokens of which 4789 (3169 harmonic and 1620 non-harmonic) (see Table 2) were used in training the model after manual segmentation in Praat[8]. The unannotated data fits the model as 1s long waveforms sampled at 16 kHz. The model's latent space contained 7 binary latent codes (2^7=128 unique lexical classes) along with 100 uniformly distributed latent variables z ($z \sim U(-1, 1)$). After training the model for 960 epochs (~44000 steps), the generated data was analyzed. We studied the mean first formant frequencies of the vowels in inputs and outputs to quantify the presence of ATR vowel harmony [9] using Praat[8], followed by regression analysis in R [10] to assess the presence of directionality.

Results and discussion: Human-like intelligible speech was generated after 800 epochs. After 960 epochs, the model generated outputs resembling the training data ([prohori], [polox], [dekhisi], [prohori]), alongside innovations ([dekhisi], [debeku], [korisuwɑ]) (see fig. 3) and variations incorporating additional or missing sounds ([iphaleo], nɔkorilu], [kɔrobe]). Notably, it fused elements from different words ([dekhisi] from [dekhisu], [kɔrobe] from [korobi]). For both training and generated data, we observe that the mean F1 value of the target vowel is much lower in the vicinity of the trigger [+ATR] vowel than that of the [-ATR] vowel (see Fig. 2). We fit the training and generated data to *linear-mixed effects* and *linear regression* models in R [10] to assess the directionality of harmony. We hypothesize that if V2 in the V1CV2 setting explains V1 better than V1 explains V2, the dataset follows regressive vowel harmony. If not, the directionality is assumed to be left-to-right We analyze the innovative items and observe that the vowels in some of them iterate over a longer domain exhibiting long-distance harmony, and the error items display non-iterative local harmony, implying that iterative harmony may indeed be myopic. The dominance of V2[i] as a trigger vowel indicates that the model can also learn the trigger feature. The statistical analysis further suggests that the grammatical outputs follow regressive directionality. Moreover, lexical learning emerges after the training (see fig. 3). We did not observe any results with the opaque vowel [ɑ] which leads us to question whether the model learns to identify vowel opacity.

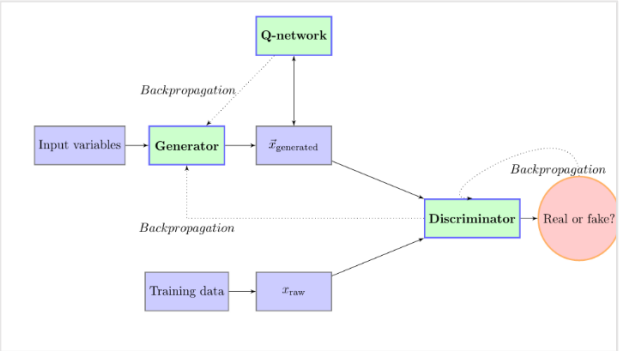


Fig. 1. Illustrative architecture of fiwGAN

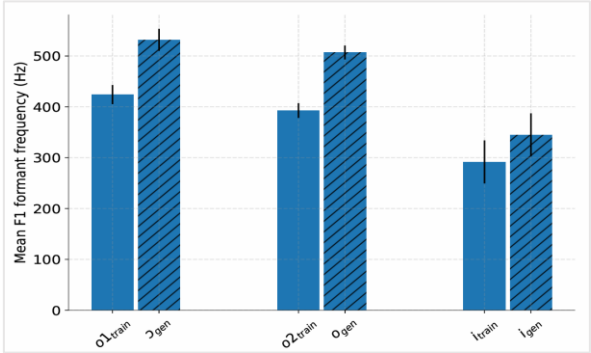


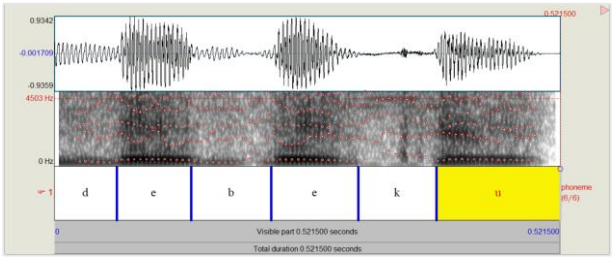
Fig. 2. F1 comparison of *[podobi]* (training data: bars) and *[pɔdobi]* (generated data: hatched bars)

Table 1. Examples of vowel harmony in Assamese

Assamese	Gloss	Suffix	Harmonized	Gloss
/pɛt/	'belly'	-u	[petu]	'pot-bellied'
/bɛpɑr/	'trade'	-i	[bepɑri]	'trader'
/zʊnɑk/	'firefly-M'	-i	[zʊnɑki]	'firefly-F'
/pɑgɔl/	'mad-M'	-i	[pɑgoli]	'mad-M'

Table 2. An exemplary training dataset

Assamese	Suffix	Harmonized
ɛlɑh	-uwɑ	elehuwɑ
ɑlɑx	-uwɑ	ɑloxuɑ
dilɛ	-i	dilei
nokorilɛ	-u	nokorileu



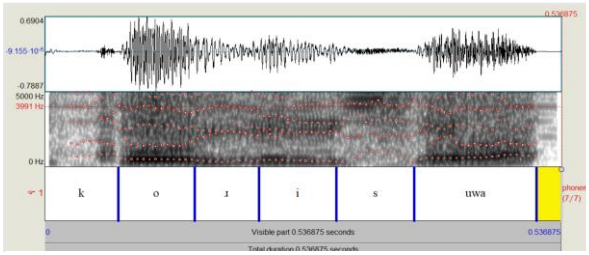


Fig. 3. Spectrograms of novel but grammatical items generated by fiwGAN: (a) '*debeku'* (left), (b) *'korisuwa'* (right), a novel but lexically meaningful word (korisuwa *'do see'*).

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.
[2] Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry, 39(3), 379-440.
[3] Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. Proceedings of the Society for Computation in Linguistics, 3(1), 149-159.
[4] Beguš, G. (2021). CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. Neural Networks, 139, 305-325.
[5] Mahanta, S. (2008). Directionality and locality in vowel harmony: with special reference to vowel harmony in Assamese (Doctoral dissertation, LOT Utrecht).
[6] Mahanta, S. (2008). Local vs. non-local consonantal intervention in vowel harmony. In Proceedings of ConSOLE XIV (Vol.165, p. 188).
[7] Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. Cognition, 171, 95-107.
[8] Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.13)
[9] Olejarczuk, P., Otero, M. A., & Baese-Berk, M. M. (2019). Acoustic correlates of anticipatory and progressive [ATR] harmony processes in Ethiopian Komo. Journal of Phonetics, 74, 18-41
[10] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.