Learning Vowel Harmony from Speech: What Emerges, What Requires Additional Structure

Sneha Ray Barman [1], **Shakuntala Mahanta**[1,2], **Neeraj Kumar Sharma** [1,3]
[1]Centre for Linguistic Science & Technology, IIT Guwahati
[2]Humanities & Social Sciences, IIT Guwahati
[3]Mehta Family School of Data Science & Artificial Intelligence, IIT Guwahati
{sneha.barman, smahanta, neerajs}@iitg.ac.in

Abstract

Vowel harmony is a non-local phonological process in which vowels within a word share a common feature value. We examine which aspects of this process are learnable from speech alone, using Assamese Advanced Tongue Root (ATR) harmony as a test case. We use a generative (fiwGAN) and a predictive model (wav2vec2) as comparative probes of what different speech-learning objectives make accessible. Both models learn ATR-related phonetic distinctions: fiwGAN latent variables control F1 differences, and wav2vec2 separates ATR categories with ∼81% accuracy. Wav2vec2 also learns global feature agreement (98% accuracy) and recognizes the blocker /ɑ/ (27.9% agreement drop in blocked contexts). However, neither model shows strong evidence for regressive spreading (d = 0.12) or iterative harmony across longer words. These results suggest that raw speech learning exposes the acoustic and co-occurrence structures, while directional phonological generalization may require morphology, lexical alternations, or temporal and architectural biases.

Keywords: phonological learning; inductive biases; vowel harmony; self-supervised learning; distributional learning

Introduction

Human infants acquire phonological knowledge from continuous speech despite variability across talkers, speaking rates, and contexts. Over time, they become sensitive to the contrasts and regularities in their native language, showing evidence for category formation and pattern learning from distributional input (Kuhl, 1991; Maye et al., 2002; Saffran & Thiessen, 2003; Werker & Tees, 1984). A core question is: which learning mechanisms and inductive biases enable learners to progress from the raw acoustic input signal to structured phonological generalizations?

The question of what phonological patterns can be recovered from distributional representations of speech, and what additional biases or learning signals support generalization beyond surface forms, sits at the center of a longstanding debate (Albright & Hayes, 2003; Hayes & Wilson, 2008; Marcus et al., 1999; Seidenberg & Elman, 1999). The discussion becomes more consequential for non-local processes that require learners to integrate information across a word and to generalize beyond memorized surface forms. Results from artificial grammar learning experiments reveal that it is harder for adults, children, and non-human animals to learn non-local dependencies than local ones (Wilson et al., 2020), while some evidence suggests that it is more challenging to learn vowel dependencies than consonant dependencies (Weyers et al., 2022). Vowel harmony, one such non-local phenomenon, requires all vowels in a word to share the same feature. As a result, a learner must recover multiple interacting properties, including the domain of harmony, the direction of harmony spreading, the trigger, and the target vowel participating in the alternation, as well as the vowels that block the spreading (Archangeli & Pulleyblank, 1989; Finley, 2010; Rose & Walker, 2011). Vowel harmony, thus, provides an ideal test case for distinguishing statistical pattern detection from rule-based phonological computation.

Evidence from infant research suggests that learners can be sensitive to harmony-like structure even without exposure to harmony languages (Götz et al., 2024; Kuhl et al., 2006; Solá-Llonch & Sundara, 2025; Sundara et al., 2018; Werker, 2024; Werker & Tees, 1984). Early research posited that infants distinguish between harmonic and disharmonic vowels due to a domain-general perceptual bias, in which listeners group sounds that share similar acoustic or articulatory features (Creel et al., 2004; Newport & Aslin, 2004; Wilson et al., 2020). Mintz et al. (2018) argued that vowel-to-vowel regularities can serve as cues for segmentation and be detected by infants in laboratory settings. Recent work confirms this sensitivity persists developmentally (Solá-Llonch & Sundara, 2025; Sundara et al., 2018). These findings demonstrate that learners may possess biases that highlight vowel-based feature dependencies. One explanation is that perceptual grouping, a domain-general tendency to group acoustically similar sounds, allows infants to detect feature agreement without representing harmony as a directional process (Creel et al., 2004; Newport & Aslin, 2004). However, full phonological knowledge may require progressing beyond this statistical sensitivity to acquire rule-like generalizations.

Earlier computational studies on vowel harmony learnability mainly focused on text data (Andersson et al., 2019; Hayes & Wilson, 2008; Mayer & Nelson, 2020), but which parts of a phonological pattern are recoverable from the speech signal itself before explicit labels for phones, features, or morphemes are supplied remains unanswered. To address this gap, we use two speech models with different learning objectives as comparative probes: a generative model trained with an adversarial loss (*featural infowaveGAN or fiwGAN*) (Beguš, 2020) and a predictive model trained with contrastive masking (*wav2vec2-large-xlsr53*) (Baevski et al., 2020; Conneau et al., 2021), and investigate whether rule-like representations

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

emerge inside the network without task-specific training on harmony. FiwGAN tests whether a generative model trained on Assamese word forms develops latent control over ATR-related acoustics. In contrast, xlsr-53 tests whether a large contextual speech model makes ATR and harmony-related information recoverable from frozen representations. These models differ in both architecture and learning objective, so we interpret their contrasting patterns as revealing what each approach achieves on the same phonological learning task.

The present paper extends Ray Barman et al. (2024), which first applied fiwGAN to Assamese vowel harmony and reported that generated vowels showed ATR-consistent F1 differences. We use the same Assamese speech corpus but ask a more specific process-level question: whether the models encode only ATR-related acoustic distinctions or also support process-level properties of harmony, including feature agreement, opacity, directionality, and iterativity. We evaluate both models on Assamese ATR harmony using a set of process-level tests. We examine (a) global feature agreement, (b) recognition of the low vowel /ɑ/ as a blocker, and (c) rules associated with phonological computation, such as directionality bias and long-distance iterativity. We ask which harmony properties emerge from distributional learning in speech under different model biases, and which properties require additional constraints or signals.

Our results reveal a dissociation informative for theories of phonological acquisition. The predictive model (wav2vec2) robustly supports global ATR agreement and sensitivity to blocking by /ɑ/, consistent with learning contextual dependencies and exceptions. However, it shows no meaningful directionality bias and no distance-based decay, suggesting that the learned representation captures harmony as global statistical co-occurrence rather than a directional, iterative process. The generative model (fiwGAN) learns phonetic categories tied to ATR but does not coordinate vowels across a word. Neither learning objective induces the rule-governed vowel harmony process, pointing to what distributional learning alone cannot recover, and what additional structure or learning signals would be required.

Model Architectures

Model 1: Featural InfoWaveGAN

Featural InfowaveGAN (fiwGAN), a fusion of WaveGAN (Donahue et al., 2019) and InfoGAN (Chen et al., 2016), was proposed by Beguš (2020). It contains three deep convolutional networks: a Generator, a Q-network, and a Discriminator. Generator, a 1D transposed convolutional neural network with five upsampling layers, takes a set of uniformly distributed latent variables (z ∼ U(-1,1)) as input (see Figure 1). The Generator's latent space comprises binary codes and latent variables. As a result, the network treats binary variables as features (ϕ) where each variable corresponds to a single feature (ϕ^n). The Generator outputs 1-second acoustic samples at 16 kHz, which, along with real data, are passed to the Discriminator, trained with a Wasserstein loss, and to the Q-network, which recovers latent codes from the output to enable lexical learning. We trained the model on the full Assamese speech dataset (4789 1-second-long audio files, each corresponding to an Assamese word token) for 1000 epochs using a batch size of 64 (other hyperparameters were the same as Beguš (2020)).

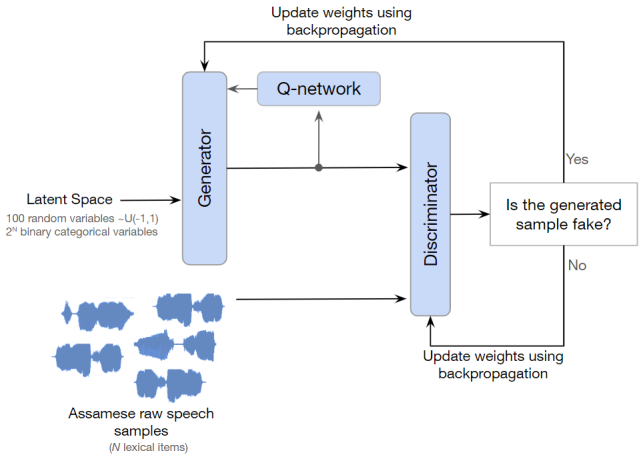

Figure 1: Architecture of Featural InfoWaveGAN (fiwGAN) displaying the Generator, the Discriminator, and the Q-network (Beguš, 2020; Ray Barman et al., 2024)

Model 2: wav2vec2-large-xlsr53

Self-supervised speech models (S3Ms), such as wav2vec2 (Baevski et al., 2020), have been remarkably successful in uncovering linguistic representations from unlabeled, raw audio data. Trained on large datasets, these models develop internal layers that often correlate with phonetic and phonological features, such as phoneme identity and manner of articulation (Cormac English et al., 2022; English et al., 2024). The wav2vec2 architecture comprises a convolutional feature encoder and a transformer encoder. A convolutional feature encoder processes raw waveforms into frame-level representations, which a 24-layer transformer then contextualizes via self-attention (see Figure 2). The model is trained with a predictive objective that forces it to learn contextual dependencies.

The contrastive loss of wav2vec2 contrasts with fiwGAN's adversarial loss, which reconstructs speech from latent variables without explicitly modeling sequential dependencies. We use wav2vec2-large-xlsr53 (Conneau et al., 2021) because it is a widely used multilingual self-supervised speech model whose representations have been shown to encode phonetic and phonological information (Choi et al., 2024; Cormac English et al., 2022; Gauthier et al., 2025; Pasad et al., 2023). Xlsr-53 was pretrained on approximately 56,000 hours of speech from 53 languages, among which approximately 55 hours of Assamese speech from the BABEL corpus is included. We use the model's frozen pretrained weights without finetuning to ensure that any phonological structure we

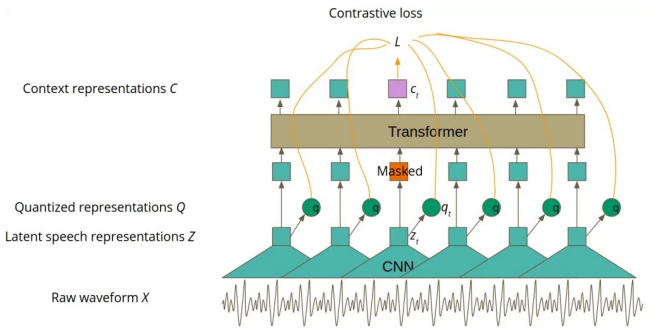

Figure 2: Architecture of wav2vec2 consisting of a convolutional feature encoder and a transformer (Baevski et al., 2020).

detect reflects representation learned through self-supervised pretraining rather than task-specific adaptation. These asymmetric training conditions reflect the models' distinct roles as probes: fiwGAN's latent space is interpretable only after training on the target data, whereas xlsr-53's frozen representations are tested on what multilingual pretraining already makes accessible.

Materials and Methods

Dataset

Our test case is Assamese advanced tongue root (ATR) harmony. Assamese is an Indo-Aryan language spoken in the state of Assam, India, by approximately 15 million speakers. The language has eight surface vowels and poses a learnability challenge due to its long-distance iterative harmony system. The [+high, +ATR] vowels [i,u] are the trigger vowels harmonizing the [-high, -ATR] vowels [ɛ, ʊ, ɔ] to [e, u, o], respectively. The [+low, -ATR] vowel /ɑ/ is the opaque vowel that blocks the right-to-left regressive harmony spreading, except for [-uwɑ] and [-ijɑ] suffixes.

Table 1: Example of ATR harmony in Assamese

Root	Gloss	Suffix	Derived	Gloss	Type
bɔx	'settle'	ɔ+ti	boxoti	'settlement'	Harmonized
lɛkʰ	'write'	i+kɑ	lekʰikɑ	'writer-F'	Harmonized
pʰɛdɛlɑ	'ugly-M'	-i	pʰedeli	'ugly-F'	Harmonized
zʊnɑk	'firefly-M'	-i	zʊnɑki	'firefly-F'	Bare (blocked by /ɑ/)
bʰut	'ghost'	-ɛ	bʰutɛ	'ghost-ERG'	Bare

We use the speech corpus introduced in Ray Barman et al. (2024), comprising 4789 tokens from 82 unique Assamese words recorded by 14 L1 Assamese speakers (8 females and 6 males, aged 19-35) from the Upper Assam region. Each speaker produced between 4 and 6 repetitions per item. The target words (X) were manually annotated and extracted from the carrier frame (moi X buli kolu) using Praat (Boersma & Weenink, 2009). Based on existing accounts of Assamese vowel harmony[1], we assigned two categories to the tokens: harmonized (all vowels in the word share the same ATR feature) and bare (words with mixed ATR vowels). This resulted in 2928 harmonized and 1861 bare word tokens, and 15451 vowel tokens ([+ATR]: 9647, [-ATR]: 5804).

[1]For details on Assamese vowel harmony, see Mahanta (2008, 2012)

Representational Analysis of fiwGAN

Ray Barman et al. (2024) showed that fiwGAN generates vowel sequences exhibiting ATR-consistent F1 differences, with [i] frequently occurring in the trigger position. The present work extends that analysis using the interpretation and probing techniques of Beguš and Zhou (2022a, 2022b) to examine which latent variables control ATR-related acoustics, and whether any learned coordination generalizes across vowels.

Latent variable identification To identify which latent variables control ATR-related acoustics, we trained a Random Forest classifier on the Discriminator's final layer activations to predict ATR categories on real training data, then applied it to annotate generated samples. Using this classifier, we generated 2000 samples from random latent vectors and extracted their predicted ATR labels. We then applied two complementary feature selection methods (the chi-square test and L1-penalized logistic regression, or LASSO) to identify which of the 100 latent variables best predict the ATR category.

Causal interpolation Correlation between a latent variable and an acoustic feature does not establish that the variable causes changes in that feature. Following the interpretability method proposed by Beguš and Zhou (2022a, 2022b), we swept each of the seven identified variables from -5 to +5, beyond their training range of [-1,1], while holding the remaining variables fixed at neutral values. This out-of-range sweep is used as a diagnostic for the direction and monotonicity of latent control rather than as a claim about the natural distribution of generated outputs; we therefore interpret the direction and consistency of the F1 change rather than its absolute magnitude. We generated an audio sample for each interpolation step and extracted f0, F1, F2, intensity, and voiced duration using Parselmouth (Jadoul et al., 2018). We computed Pearson correlations between latent variables and acoustic measures to assess whether each variable systematically controls formant trajectories.

Process-level tests To test whether harmony spreads across vowels, we generated CVCVC sequences by concatenating separately generated syllables. We manipulated one vowel's ATR-encoding latent variable while holding the other vowel's corresponding latent at a neutral value. Then, we measured whether the manipulated vowel's F1 change resulted in changes in the non-adjacent vowel. To test directionality, we compared F1 slopes when the manipulated vowel appeared in word-initial versus word-final position to assess whether the model shows a bias toward regressive or progressive spreading.

Representational Analysis of wav2vec2

Representation extraction and best layer identification We extracted layerwise representations from wav2vec2-large-xlsr53 by forward-passing each waveform through the frozen representation and mean-pooling hidden states over each vowel's timespan. The target words were manually segmented, and vowel boundaries were detected using an energy-based algorithm and validated against the phonetic transcriptions performed in Praat. To identify the layer at which ATR information is most accessible under contextual prediction, we trained logistic regression probes on each of the 24 transformer layers to predict the target vowel's ATR category from contextual representations while masking the target vowel's own frames before pooling. The layer selected by this procedure was used for all subsequent process-level tests.

Process-level tests We designed four tests to examine whether the model learns properties of phonological processes beyond featural co-occurrence. The cross-vowel masking test assesses feature agreement by predicting each vowel's ATR category from the remaining vowels in the word. To mask the target vowel, we set its time-aligned frames to zero before pooling. We report three metrics: (i) cross-vowel prediction accuracy: the proportion of correctly predicted ATR categories when the target vowel is masked, (ii) internal consistency: the proportion of words where all vowels receive identical ATR predictions, and (iii) word-level agreement: the proportion of words where the predicted ATR category matches the ground-truth harmony label. The asymmetric masking test compares prediction accuracy when using only the right context (regressive) to only the left context (progressive) to assess directional bias. The opacity test compares ATR agreement between vowel pairs separated by /ɑ/ (blocked, mean distance: 2.29 vowels) and vowel pairs separated by non-/ɑ/ vowels at comparable distance (unblocked, mean: 2.38 vowels). The distance decay test measures contextual prediction and internal consistency as a function of trigger-target distance to assess long-distance iterativity.

Results

The reported findings address two primary questions: (i) whether both models learn ATR as a phonetic distinction, and (ii) whether this encoding reflects grammatical rule learning or statistical co-occurrence. We evaluate the second question through a set of process-level tests targeting universal properties of vowel harmony (feature agreement, opacity, directionality, and long-distance iterativity). A methodological asymmetry structures the results. Wav2vec2 produces token-aligned contextual vowel embeddings that support masking-based tests, whereas fiwGAN's generative architecture lacks such a mechanism. Feature agreement and opacity tests are therefore applied only to wav2vec2, and fiwGAN is evaluated via latent interpolation and spreading tests on the generated outputs. We first establish that both models encode ATR-related phonetic distinctions, then test constraint-like properties (feature agreement and opacity), and finally assess rule-like properties (directionality and iterativity).

ATR Contrast Encoding

FiwGAN The chi-square feature selection and LASSO regression tests converged on 7 latent variables that encode ATR−related acoustics (Table 2). Variables z5, z6, and z7 showed the strongest associations ($\chi^2 > 7.5$, $p < 0.006$). A logistic regression classifier trained only on these 7 variables achieved 88.4% accuracy in predicting ATR labels, retaining 97.9% of the performance of the complete 100-variable model. To verify whether these identified variables systematically control the acoustic distinctions of the ATR vowels, we compared F1 values of the generated samples annotated as [+ATR] and [-ATR]. Generated outputs showed an F1 difference of 144.4 Hz (Cohen's $d = -1.133$), with [+ATR] vowels lower in F1 than [-ATR]. Permutation testing with 1000 iterations confirmed that the observed difference was significantly above chance (permutation $p < 0.0001$). A lower F1 value for [+ATR] vowels than for their [-ATR] counterparts is a well-established acoustic distinction in the literature (Hess, 1992; Olejarczuk et al., 2019; Starwalt, 2008), and is notable here as the model received no explicit phonological labels during training.

Table 2: Top 7 z-variables controlling ATR-encoding in fiwGAN's Generator

z-variable	chi-square (χ^2)	p-value	Lasso (β) coefficient
z7	9.17	0.0025	-0.5320
z5	8.53	0.0035	-0.4015
z6	7.54	0.0060	-0.3549
z26	7.14	0.0075	0.4100
z4	6.99	0.0082	-0.3648
z60	5.98	0.0145	-0.3516
z81	5.84	0.0156	0.3662

Latent interpolation revealed that all 7 variables exert a consistent directional influence on the F1 trajectory. The Pearson correlations ranged from $|r| = 0.712$ to 0.950 (all $p < 0.001$). When interpolated from -5 to +5, each variable produced systematic F1 changes ranging from 302.5 Hz (z26) to 911.3 Hz (z60). Variable z7 emerged as the primary controller of vowel quality, producing the largest ranges for F1 (787 Hz, r=0.966) and F2 (518 Hz, $r = 0.946$)(see Figure 3).

Wav2vec2 ATR classification accuracy improved steadily through lower and mid transformer layers. The classifier's performance peaked at Layer 12, with an accuracy of 81.1% ±8.3%, macro-F1 of 0.787 ± 0.081, and ROC-AUC of 0.891 ± 0.072 (Figure 4). Performance declined in deeper layers, suggesting ATR-relevant information becomes less linearly separable as representations encode higher-level contextual dependencies. For all subsequent process-level tests, we used Layer 12 representations, which achieved 93.3% accuracy in

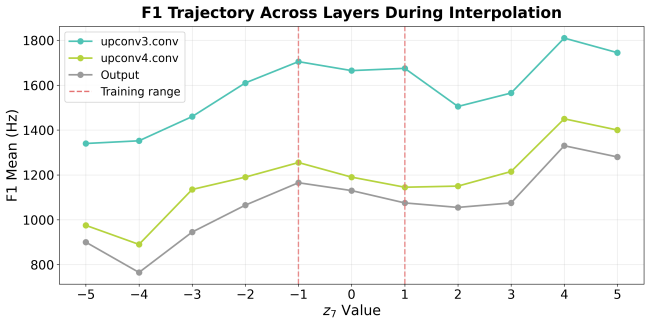

Figure 3: F1 mean across fiwGAN Generator layers during z_7 interpolation from -5 to +5. Dashed lines indicate training range boundaries (-1, +1). All layers show systematic F1 control.

contextual ATR prediction (predicting a vowel's ATR category from neighboring vowels while masking the target)[2]. The higher contextual prediction accuracy (93.3% vs. 81.1%) indicates that ATR is highly predictable from local context, even when the target vowel's own acoustic information is unavailable.

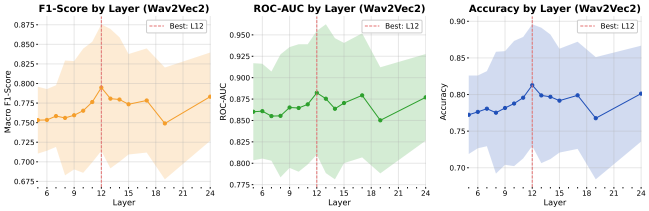

Figure 4: Layerwise ATR classification performance in wav2vec2. All three metrics (F1-score, ROC-AUC, and accuracy) peak at Layer 12. Shaded regions indicate cross-validation standard deviation.

Feature Agreement and Opacity Learning

Feature agreement The cross-vowel masking probe predicted each vowel's ATR category using only contextual representations from the remaining vowels in the word. Across 15451 vowel tokens, cross-vowel prediction accuracy reached 98% (t=466.75, p <0.001), word-level agreement (whether the consistently predicted ATR category matched the ground-truth label) reached 99.5%, and internal consistency (whether all vowels in a word received the same predicted ATR category) reached 98.5%. Cohen's d effect sizes ranged from 3.4 to 6.8 (very large effect)[3]. These near-ceiling values indicate that wav2vec2's intermediate representations encode strong within-word ATR co-occurrence, a core property of vowel harmony. However, they may also reflect per-phoneme identity encoding, disentangling which would require nonword generalization tests.

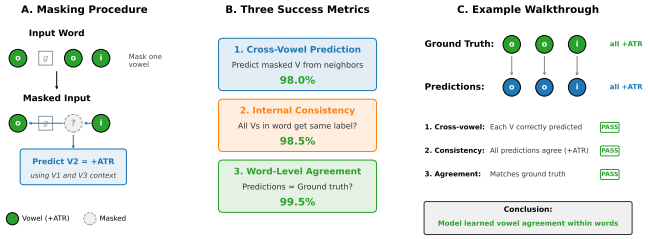

Figure 5: Schematic of the cross-vowel masking test: one vowel is masked, and its ATR label is predicted from the remaining vowels' contextual representations.

We further examined how masking individual vowels affects predictions. At the vowel level, only 12.5% of other vowels changed their predicted ATR category when one vowel was masked. However, word-level harmony classification (harmonized vs. bare) resulted in a 21.6% accuracy drop when any single vowel is masked. Masking the rightmost vowel position caused the largest accuracy drop (28.3%). This is consistent with some reliance on the rightmost position, which corresponds to the trigger in regressive harmony, though positional frequency differences in the corpus cannot be ruled out.

Opacity To test whether wav2vec2 recognizes /ɑ/ as a harmony blocker, we compared ATR agreement between two conditions: blocked (vowel pairs intervened by /ɑ/) and unblocked (pairs at comparable distance without intervening /ɑ/). Vowel pairs separated by /ɑ/ agreed in predicted ATR only 53.5% of the time compared to 81.4% for the unblocked pairs. The blocking effect was +27.9% (t=16.2, p <0.0001, Cohen's d=1.42). This large effect indicates that the model less accurately predicts ATR agreement when /ɑ/ intervenes between vowels, consistent with the Assamese ATR harmony pattern, in which /ɑ/ blocks spreading.

Directionality

FiwGAN We swept an ATR-encoding latent for either the preceding or following vowel while holding a neutral mid-vowel constant. If the model had internalized regressive harmony, we would expect larger effects when the manipulated vowel follows the target. We observed no significant difference in F1 slope or mean shift between leftward and rightward conditions (p >0.3). This direction-independent behavior indicates that the Generator lacks a bias toward regressive harmony despite being trained on a dataset with strictly regressive patterns.

Wav2vec2 We compared prediction accuracy when using only right context (regressive: predicting left vowels from

[2]Layerwise probing measures ATR decodability from vowel representations, whereas contextual ATR prediction measures predictability from context with the target vowel masked; these are distinct evaluations.

[3]Cohen's d standardized effect sizes: small effect $d \approx 0.2$; medium effect $d \approx 0.5$; large effect $d \approx 0.8$ or greater.

right context) versus only left context (progressive: predicting right vowels from left context). The model predicted left vowels' ATR category from right context with 72.3% accuracy and right vowels' ATR from left context with 67.7% accuracy. This resulted in a directional bias of +4.7% ($t = 8.94, p < 0.001$, Cohen's d =0.12). The statistically significant but negligible effect size suggests that the model shows no meaningful preference for regressive spreading and does not strongly encode the directional mechanism of Assamese vowel harmony.

Cross-vowel Coordination and Distance Decay

FiwGAN: cross-vowel coordination We generated (C)VCVCV(C) sequences by concatenating separately generated syllables, manipulated one vowel's ATR-encoding latent while holding the other vowels' latent at a neutral value, and measured whether this produced F1 changes in the non-adjacent vowel. Given independent generation, the absence of cross-vowel effects indicates that the model's learned ATR control remains local rather than reflecting an emergent spreading dependency. If the model had learned iterative harmony, the manipulated vowel should systematically affect the non-adjacent vowel's acoustic properties (especially F1). Instead, the non-local vowel's F1 remained flat across the first vowel's sweep, and vice versa. Slopes of F1 change were identical irrespective of vowel position.

Wav2vec2: distance decay Distance decay was tested by grouping trigger–target pairs by the number of intervening vowels and measuring mean target-prediction accuracy as a function of distance. Individual vowel prediction accuracy remained high and stable across distances 1 to 4 (0.976 – 0.986), with no significant correlation between distance and accuracy (r=0.020, p=0.17). However, internal consistency, the proportion of words where all vowels received identical ATR predictions, dropped sharply across word-length from 74.6% (2-vowel words) to 61.4% (3-vowel words) to 41.3% (4+ vowel words). Therefore, the highly accurate pairwise predictions mask the model's struggle to assign consistent ATR values across all vowels in longer words, thereby preventing it from implementing an iterative mechanism that enforces harmony across all positions.

Discussion

In this paper, we examined which aspects of vowel harmony can be recovered from distributional statistics in raw speech. We used two speech models with different learning objectives (fiwGAN and wav2vec2) as computational probes to learn vowel harmony from acoustic signals. Across both models and all process-level tests, the results point in the same direction. Both models learn the phonetic distinctions between ATR categories. Wav2vec2 goes further, additionally learning global feature agreement and harmony blocking by /ɑ/. These observations reflect the scope and limits of distributional learning for phonological acquisition.

The results on feature agreement and opacity for wav2vec2 align with perceptual-grouping accounts (Creel et al., 2004; Newport & Aslin, 2004). Infants can discriminate between harmonic and disharmonic sequences even without prior exposure to a harmony language (Altan et al., 2016; Solá-Llonch & Sundara, 2025), and our models appear to reach a similar sensitivity through distributional patterns in the acoustic input. However, our models do not progress beyond this point. The opacity finding, where ATR agreement dropped by 27.9% when /ɑ/ intervened between the distance-matched vowel pairs, suggests that something about the distributional pattern of blocking is recoverable from surface statistics. However, what exactly the model is doing is less clear. It may have learned /ɑ/ as a categorical boundary, or it may be picking up on coarticulatory or prosodic cues that tend to co-occur with /ɑ/ in this corpus. The near-ceiling accuracy values for global feature agreement are consistent with the model encoding ATR as a shared feature, but a classifier that learns separate representations for individual vowels would produce similar numbers without representing ATR abstractly at all. Whether these results reflect genuine feature agreement or phoneme-specific encoding cannot be determined from the current data alone. Distinguishing between these probabilities would require testing on acoustically matched contexts or novel word forms.

The directionality and iterativity tests put forth interesting observations. Wav2vec2 showed a statistically significant but negligible directional bias (d = 0.12), whereas fiwGAN showed none at all, despite being trained on a dataset with strictly regressive spreading. The internal consistency results clarify this. The pairwise vowel prediction stayed high and stable across distances, but word-level consistency fell from 74.6% in 2-vowel words to 41.3% in words with 4 or more vowels. The drop suggests the predictions are driven by local co-occurrence rather than a global feature specification. The fact that F1 slopes were identical in fiwGAN regardless of which vowel was manipulated indicates that each vowel's quality is determined by its own latent variable with no detectable influence on its neighbors. Both models, then, appear to have learned the surface statistics of harmony without learning the mechanism behind it.

Our overall findings suggest that full phonological learning requires combining distributional mechanisms with structured representations. Human adults pick up directional biases and generalize to new forms after limited exposure (Finley & Badecker, 2009, 2012), and the gap between what they do and what these models do points to learning signals that acoustics alone cannot provide. Morphological alternations, since root-affix paradigms in Assamese make spreading directionality visible in a way surface statistics do not (Albright & Hayes, 2003), and architectural asymmetry, such as a strictly unidirectional context window, are some of them. Future work should extend this framework cross-linguistically across harmony systems with different structural properties.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161. https://doi.org/10.1016/S0010-0277(03)00146-X

Altan, A., Kaya, U., & Hohenberger, A. (2016). Sensitivity of turkish infants to vowel harmony in stem- suffix sequences: Preference shift from familiarity to novelty. *BUCLD 40 Online Proceedings Supplements.*

Andersson, S., Dolatian, H., & Hao, Y. (2019). Computing vowel harmony: The generative capacity of search & copy. *Proceedings of the Annual Meetings on Phonology*.

Archangeli, D., & Pulleyblank, D. (1989). Yoruba vowel harmony. *Linguistic inquiry*, *20*(2), 173–217.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Beguš, G. (2020). Ciwgan and fiwgan: Modeling lexical learning from raw acoustic data in generative adversarial phonology.

Beguš, G., & Zhou, A. (2022a). Interpreting intermediate convolutional layers in unsupervised acoustic word classification. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8207–8211. https://doi.org/10.1109/ICASSP43922.2022.9746849

Beguš, G., & Zhou, A. (2022b). Interpreting intermediate convolutional layers of generative CNNs trained on waveforms. *ieee$_{ja}$slp*, *30*, 3214–3229.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1.13). http://www.praat.org

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, *29*.

Choi, K., Pasad, A., Nakamura, T., Fukayama, S., Livescu, K., & Watanabe, S. (2024). Self-supervised speech representations are more phonetic than semantic. *Proc. Interspeech 2024*, 4578–4582. https://doi.org/10.21437/Interspeech.2024-1157

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of Interspeech*.

Cormac English, P., Kelleher, J. D., & Carson-Berndsen, J. (2022, July). Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. In G. Nicolai & E. Chodroff (Eds.), *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 83–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.9

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(5), 1119–1130. https://doi.org/10.1037/0278-7393.30.5.1119

Donahue, C., McAuley, J., & Puckette, M. (2019). Adversarial audio synthesis. *ICLR*.

English, P., Kelleher, J., & Carson-Berndsen, J. (2024). Searching for structure: Appraising the organisation of speech features in wav2vec 2.0 embeddings. *Proc. Interspeech 2024*, 4613–4617.

Finley, S. (2010). Exceptions in vowel harmony are local. *Lingua*, *120*(6), 1549–1566. https://doi.org/10.1016/j.lingua.2009.10.003

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*(3), 423–437. https://doi.org/10.1016/j.jml.2009.05.002

Finley, S., & Badecker, W. (2012). Learning biases for vowel height harmony. *Journal of Cognitive Science*, *13*(3), 287–327. https://doi.org/10.17791/JCS.2012.13.3.287

Gauthier, J., Breiss, C., Leonard, M. K., & Chang, E. F. (2025, November). Emergent morpho-phonological representations in self-supervised speech models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 28067–28086). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1425

Götz, A., Krasotkina, A., Schwarzer, G., & Höhle, B. (2024). Asymmetries in infants' vowel perception: Changes in vowel discrimination in german learning 6- and 9-month-old infants. *Language and Speech*, *67*(4), 1135–1149. https://doi.org/10.1177/00238309241228237

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Hess, S. (1992). Assimilatory effects in a vowel harmony system: Akan. *J. Phonetics*, *20*(4), 417–435.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/https://doi.org/10.1016/j.wocn.2018.07.001

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Mahanta, S. (2008). *Directionality and locality in vowel harmony: With special reference to vowel harmony in assamese*. Netherlands Graduate School of Linguistics.

Mahanta, S. (2012). Assamese. *Journal of the International Phonetic Association*, *42*(2), 217–224.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283 5398*, 77–80. https://api.semanticscholar.org/CorpusID:6261323

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3

Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. *Society for Computation in Linguistics*, *3*(1).

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. https://doi.org/10.1016/j.cognition.2017.10.020

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162. https://doi.org/10.1016/S0010-0285(03)00128-2

Olejarczuk, P., Otero, M. A., & Baese-Berk, M. M. (2019). Acoustic correlates of anticipatory and progressive [ATR] harmony processes in Ethiopian Komo. *J. Phonetics*, *74*, 18–41.

Pasad, A., Shi, B., & Livescu, K. (2023). Comparative layer-wise analysis of self-supervised speech models. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Ray Barman, S., Mahanta, S., & Sharma, N. (2024). Deciphering assamese vowel harmony with featural infowavegan. *Proc. Interspeech 2024*, 1510–1514. https://doi.org/doi:10.21437/Interspeech.2024-1990

Rose, S., & Walker, R. (2011). Harmony systems. *The handbook of phonological theory*, 240–290.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494. https://doi.org/10.1037/0012-1649.39.3.484

Seidenberg, M. S., & Elman, J. L. (1999). Networks are not 'hidden rules'. *Trends in Cognitive Sciences*, *3*(8), 288–289. https://doi.org/10.1016/S1364-6613(99)01355-8

Solá-Llonch, E., & Sundara, M. (2025). Young infants' sensitivity to precursors of vowel harmony is independent of language experience. *Infant Behavior and Development*, *78*, 102032. https://doi.org/10.1016/j.infbeh.2025.102032

Starwalt, C. F. (2008). *The acoustic correlates of ATR harmony in nine vowel harmony systems: A cross-linguistic study* [Ph.D. Dissertation]. University of Illinois at Urbana-Champaign.

Sundara, M., Ngon, C., Skoruppa, K., Feldman, N. H., Onario, G. M., Morgan, J. L., & Peperkamp, S. (2018). Young infants' discrimination of subtle phonetic contrasts. *Cognition*, *178*, 57–66. https://doi.org/10.1016/j.cognition.2018.05.009

Werker, J. F. (2024). Phonetic perceptual reorganization across the first year of life: Looking back. *Infant Behavior and Development*, *75*, 101935. https://doi.org/10.1016/j.infbeh.2024.101935

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Weyers, I., Männel, C., & Mueller, J. L. (2022). Constraints on infants' ability to extract non-adjacent dependencies from vowels and consonants. *Developmental Cognitive Neuroscience*, *57*, 101149. https://doi.org/10.1016/j.dcn.2022.101149

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., van der Kant, A., Smith, K., & Rey, A. (2020). Non-adjacent dependency learning in humans and other animals [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12381]. *Topics in Cognitive Science*, *12*(3), 843–858. https://doi.org/10.1111/tops.12381