Do Self-supervised Models Learn Phonological Rules? Evidence from Assamese ATR Harmony

Anonymous ACL submission

Abstract

Self-supervised speech models (S3Ms) often achieve high accuracy on phonological probing tasks; however, this accuracy may not accurately reflect the acquisition of grammatical processes. In this paper, we present an interpretive analysis of wav2vec2.0's representations using Assamese Advanced Tongue Root (ATR) vowel harmony as a case study. Instead of treating probing accuracy as evidence of rule learning, we combine layerwise probing with a set of masking-based tests designed to distinguish between global feature agreement and structure-phonological computation. Comparing two multilingual wav2vec2.0 variants, we show that ATR features are linearly decodable in the intermediate layers (peaking at ∼80% accuracy) and that models successfully encode within-word feature agreement and sensitivity to an opaque vowel (demonstrating strong blocking effects with ∼30% accuracy drop). At the same time, these representations provide limited evidence for rule-governed properties, such as directionality and trigger specificity. This raises an important interpretability question: do models pass individual phonological tests without implementing a systematic generative process? Our overall observations demonstrate that high probing accuracy and task-specific masking tests can sometimes overstate grammatical competence. We argue that phonological processes provide a valuable benchmark for interpretability methods, highlighting the importance of evaluating constraint interactions rather than isolating properties when analyzing neural speech models.

1 Introduction

Self-supervised speech models (S3M) have achieved remarkable success in uncovering linguistic representations from raw audio data (Jaiswal et al., 2020; Manning et al., 2020; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). Trained on large unlabeled datasets, these models develop internal layers that often correlate with phonetic and phonological features, such as phoneme identity and manner of articulation (English et al., 2022, 2023). However, high classification accuracy on a phonological contrast does not guarantee that a model has learned the grammatical rules controlling the contrast. The present research, therefore, emphasizes the central question of explainability in neural speech models: what do these models learn to utilize the features they encode?

To address this, we use vowel harmony as a benchmark for evaluating grammatical competence in S3Ms. Unlike local processes such as assimilation or aspiration, which can be learned from adjacent segment co-occurrences, vowel harmony poses sensitivity to multiple interacting constraints, such as, directionality, trigger specificity, opacity, and iterativity. A model that has truly learned harmony must encode all of these rules; success on one constraint would not guarantee success on others. Assamese, an Indo-Aryan language spoken by approximately 15 million native speakers in the Northeastern region of India, shows an interesting case of long-distance harmony whereby (i) a rightmost [+high,+ATR] vowel triggers harmony on all preceding [-high, -ATR] vowels, (ii) the [+low, -ATR] vowel /ɑ/ blocks the spreading, and (iii) harmony applies iteratively, spanning multiple syllables (Mahanta, 2007, 2008).

We use the *wav2vec2-large-xlsr-53* (Conneau et al., 2021) and *wav2vec2-xls-r-300m* (Babu et al., 2022) models as frozen feature extractors and combine layerwise probing with an array of process-level tests. Our goal is to determine whether the model learns ATR harmony as a grammatical process or merely as featural co-occurrences. Understanding how and where the ATR-related features are internally encoded is also expected to inform us about how to use these self-supervised speech models for downstream tasks in language. We ask two questions through our work:

- **RQ1:** Where in the model's layers is ATR information encoded, and do the two models differ in their encoding hierarchy?

- **RQ2:** Does the model learn phonological properties of ATR harmony, such as feature agreement, directionality, opacity, and long-distance spreading, or does it only encode statistical feature co-occurrence?

We compare two multilingual wav2vec2 models: *large-xlsr-53* (pretrained on 53 languages) and *xls-r-300m* (pretrained on 128 languages). Both have similar architectures (∼300M parameters) and include Indo-Aryan languages such as Bengali in their pretraining data, enabling cross-linguistic transfer to Assamese. Comparing these models tests whether increased multilingual exposure improves grammatical learning of harmony. If both models exhibit similar limitations in capturing ATR harmony rules, this suggests that the issue is architectural or objective-based rather than data-related, which has implications for the design of self-supervised speech models.

Our findings reveal that both models achieve peak ATR classification in mid-layers (L12 for *xlsr-53*, L16 for *xls-r-300m*), reaching $\sim$ 80% accuracy for both models, but differ in layerwise encoding patterns. Our process-level tests show that both models successfully encode ATR harmony as a global feature agreement and recognize /ɑ/ as an opaque vowel that blocks harmony. However, both exhibit only weak directional bias of ∼5% and show no distance decay in harmony spreading. These results suggest that the S3M treats harmony as a global feature co-occurrence rather than directional spreading, preferring some grammatical constraints governing vowel harmony while missing others.

2 Related Work

Early works on wav2vec2 demonstrated that contrastive predictive coding over masked speech segments yields representations that capture both local acoustic structure and higher-level linguistic information (Baevski et al., 2020; Conneau et al., 2021). The xlsr-53 model, pretrained on a multilingual dataset, enhances phone recognition and cross-linguistic generalization, particularly for low-resource languages (Conneau et al., 2021).

Subsequent studies on layerwise representations of the self-supervised models have indicated the development of a representational hierarchy. Pasad et al. (2021, 2024) used Projection-weighted Canonical Correlation Analysis (PWCCA) to report that intermediate layers of the S3M best encode phonetic and phonological information, with early layers encoding spectral properties and deeper layers more abstract features. English et al. (2022, 2023) further strengthened this observation by demonstrating that phonological feature decoding (e.g., voicing, place, nasality) peaks at the middle layers. English et al. (2024) extended the earlier study on transition regions and coarticulation domains to demonstrate that S3M representations effectively capture context-sensitive phonetic details. Related studies have analyzed segmental properties, such as aspiration (Martin et al., 2023), vowel front-back distinctions (Cristofaro et al., 2025), and emergent morphophonological patterns (Gauthier et al., 2025), consistently observing that these features are linearly separable in the intermediate representations.

Evidence from research that asked whether these representations support process-level knowledge (sensitivity to context-informed alternations) is mixed. Choi et al.'s (2024) observations comparing various S3Ms revealed that their layerwise representations encode phonetic information more robustly than semantic information. English et al. (2024) pointed out their ability to learn coarticulation and assimilation as statistical tendencies rather than explicit, context-conditioned rules. Gauthier et al. (2025) examined English plural allomorphs and demonstrated that these models encode the relevant distinctions as a distributed geometric structure but exhibit limited rule-like generalization to novel forms. These findings encourage us to look at a broader perspective: when an S3M achieves high accuracy in predicting a phonological contrast, does it account for the underlying grammatical constraints or rely solely on the surface distribution?

Vowel harmony, however, remains relatively understudied in the context of S3M interpretability. Earlier literature primarily evaluated segment-level phonetic features and local coarticulatory patterns without testing whether models exhibit rule-like behavior characteristics. (Gopinath and Rodriguez, 2024) recently probed whether Turkish-trained and English-trained models encode vowel harmony differently and observed little difference in attention patterns despite Turkish's long-distance harmony. This finding, along with the existing gap in the literature, further motivates us to investigate whether

S3Ms capture harmony as a grammatical process or merely as statistical patterns.

3 Experimental Setup

3.1 Dataset

Speech data were collected from 14 native Assamese speakers (6 males, 8 females, ages $19-35$) in a soundproof booth. All participants were native speakers of the Upper Assam variety of colloquial Assamese and had no reported history of speech or hearing disorders. Recordings were made using a Tascam DR-100 MKII digital recorder at a sampling rate of 48 kHz with 16-bit resolution. The elicitation materials consisted of 82 unique Assamese words (40 displaying harmonic alternations, 42 showing no harmony) embedded in the carrier phrase *moi X buli kolu* ('I say X'). Each speaker produced each target word at least four times, yielding approximately $4,920$ potential tokens.

Stem	suffix	Surface	Harmony type
dilɛ	-i	dilei	harmonized
nokorilɛ	-u	nokorileu	harmonized
gɔrɔm	-ɔ	gɔrɔmɔt	harmonized
bɛpɑr	-i	bɛpɑri	bare (harmony blocked by /ɑ/)

Table 1: Example words from the created dataset showing harmonized and bare word types in the Assamese language.

Each of the 82 words was manually labelled as *harmonized* (all vowels share the same ATR feature) or *bare* (mixed ATR features) based on the established phonological descriptions of Assamese (Mahanta, 2007, 2008, 2012). Individual vowels were given ATR labels ([+ATR] or [-ATR]) using a manually created dictionary mapping each word to its respective IPA transcription. All recordings were downsampled to 16 kHz mono-channel for subsequent data analysis. After excluding tokens with recording errors and background noise, our final dataset contained spoken audio recordings of 4789 word tokens (harmonized:2928,bare:1861) comprising 15451 vowel tokens ([+ATR]: 9647, [-ATR]: 5804). See Table 1 for a few example words from the dataset.

3.2 Embedding extraction from Pretrained Models

For our experiment, we used the pre-trained wav2vec2 models ((Conneau et al., 2021; Babu et al., 2022)): wav2vec2-xlsr-53 and wav2vec2-xls-r-300m. The models' weights are kept frozen; we do not fine-tune them for our task to ensure the encoded information is a result of the pretraining. Both models consist of a feature encoder layer and 24 transformer layers. We forward-pass the waveform corresponding to each through the models and extract layerwise embeddings for each of the transformer layers. To extract word-level representations, we employed **mean word pooling** across the temporal dimension of the waveform corresponding to the whole word (Pasad et al., 2021, 2024). Thus, for each input word token instance, we obtain a 1024-dimensional vector per word for each of the 24 layers, from each of the models. For vowel-level analysis, we detect vowel boundaries using an energy-based algorithm and extract per-vowel representations by mean-pooling hidden states over each vowel's time span. Detected boundaries were further validated against expected vowel counts from the phonetic transcriptions. These extracted representations enable us to analyze where in the network the ATR information is most prevalent.

3.3 Probing the Models for ATR Classification

We set up a probing task to assess which layer performs the best in classifying the ATR feature in both models. To test the linear separability of the ATR feature in the embedding space of each layer, we fit a logistic regression classifier to the extracted embeddings from both models. For word-level probing, the binary classifier has to classify the 1024-dimensional embedding vector into harmonized vs bare. We performed 5-fold GroupK-Fold cross-validation with root words as groups to ensure that all tokens of a given word (lexical) item appear in either the train or the test split but not both. This evaluates the generalization across lexical items rather than memorizing specific words. Our evaluation metrics include classification accuracy and the macro-F1 score. Accuracy is reported on the held-out folds, and we aggregate the results to obtain the mean accuracy and standard deviation for each layer's probe.

For vowel-level probing, the same logistic regression classifier predicts ATR category ([+ATR] or [-ATR]) for individual vowels with GroupK-

Fold cross-validation. By comparing the logistic regression results, we can identify which model performs better and where in its layers the performance peaks. We assume that if a specific layer's representation effectively separates ATR harmony categories, the classifier will achieve high accuracy; otherwise, accuracy will be near or below the chance level (∼50%). Thus, the probing task quantifies the encoding capacity of each layer's embedding for ATR distinction.

3.4 Phonological Process Learning Testing of ATR Harmony

So far, we have tested where and how strongly ATR-related features emerge inside the model. Now, we probe the representations further to examine whether the model learns phonological aspects of advanced tongue root vowel harmony or if the results reflect an instance of featural co-occurrence. For this, we combined layerwise probing with directional and contextual manipulation. We tested directionality, trigger specificity, opacity, and long-distance harmony learning. We extract vowel-level representations and train a logistic regression probe on each of the 24 transformer layers. The probe was tasked with predicting the ATR category of a target vowel using only contextual representations from its neighboring vowels, while the target's own representations were masked. This classification task helped us identify the best-performing layer: L12 for xlsr-53 and L16 for xls-r-300m (both reached 93.3% accuracy in contextual ATR prediction)), and its representations were used for the process-learning tests. [1]

3.4.1 Cross-vowel Masking Test for Feature Agreement

We conducted the cross-vowel masking test to determine whether the model could assign the same ATR categories to all vowels within a word. Our masking algorithm iterated through each word token, masked one vowel at a time (replacing it with zeros), and attempted to predict its ATR category based on the representations of remaining vowels. We set three metrics to measure the success of the masking experiment. One of them was the above-mentioned cross-vowel prediction test. Secondly, we measured internal consistency by checking if all predicted vowels within a word shared the same ATR value. Word-level agreement, the third metric, was calculated by comparing this consistently predicted ATR category to the ground-truth ATR label of the entire word. High accuracy indicates the model has learned within-word ATR dependencies.

3.4.2 Asymmetric Masking Test for Directionality

Assamese harmony is regressive: rightmost [+ATR] vowels trigger harmony on preceding vowels. The asymmetric masking test investigates which direction the model prefers when predicting harmony. According to simple masking test predictions, masking the rightmost context should block the left vowel's ATR category prediction, while masking the left context should block the right vowel's ATR feature prediction. If the model learned regressive harmony, it should fail to predict the leftward vowels' ATR if the rightmost context is blocked. To test which directionality is preferred, we employed two types of masking tests on the representations of the best-performing layer.

L-from-R Prediction (regressive directionality): This task is designed to predict the ATR feature of the left vowel from the right context. For each target vowel, we mask its own representation and predict its ATR class (+ATR/-ATR) based on the embeddings of vowels to its right, including the rightmost potential trigger. If the model encodes regressive harmony, the right context should carry sufficient information to predict the left vowels' ATR features.

R-from-L Prediction (progressive directionality): This task predicts the ATR feature of the right vowel from the left context. The leftmost vowels' embeddings are masked, and the ATR feature of the rightward vowels must be predicted. This test determines whether the left context conveys harmony information, indicating progressive harmony.

For each word token, we computed prediction accuracy in both prediction tests We report directional bias, defined as the mean difference between the accuracy of predicting left-from-right and right-from-left contexts A positive value will indicate a preference for regressive harmony We also performed a paired t-test to assess the accuracy difference It is important to note that we ensure that the model compares the predicted ATR value with the ground-truth ATR labels for the target vowels

[1] Although the word-level harmony classifier and the contextual ATR classifier peak at the same layer in our data, they are distinct tasks. We selected layers for process testing only based on the contextual probe.

Therefore, the asymmetric masking test thoroughly examines whether context alone carries the harmony signal and which direction is preferred.

3.4.3 Trigger Specificity Test

In Assamese, [+high, +ATR] vowels function as harmony triggers To test whether the model differentiates trigger types, we split the word tokens by the ATR feature of the rightmost vowel As a result, we got two sets of candidates: one set with rightmost [+ATR] vowels and another with rightmost [-ATR] vowels For each word, we computed prediction accuracy in both regressive and progressive directions and evaluated them against the ground-truth ATR labels of the targets We compare regressive prediction accuracy between two groups High accuracy for [+ATR] triggers would indicate the model has rightly identified the trigger vowels of Assamese ATR harmony.

3.4.4 Opacity Test

The low vowel /ɑ/ blocks harmony spreading in Assamese: a [+ATR] trigger cannot harmonize vowels across an intervening /ɑ/ except for [-uwɑ] or [-ijɑ] suffixes. To test whether the models encode the harmony blocking constraint, we evaluate the ATR prediction accuracy across two conditions: the blocked condition, where the trigger and the target vowels are separated by /ɑ/ (e.g., V1 and V3 when V2 = /ɑ/, and the unblocked condition, where the target and trigger vowels are either adjacent or separated by a non-/ɑ/ vowel. An agreement is defined as both vowels receiving the same predicted ATR. If the model shows opacity-sensitive patterns, agreement should be significantly lower in blocked contexts.

3.4.5 Distance Decay Test

We conducted the distance-decay analysis to examine whether harmony strength weakens as the distance between the trigger and the target vowel increases. We expect that prediction accuracy and confidence should gradually decrease at greater distances from the trigger. For three or more vowels (n = 3788), we defined distance as the number of intervening vowels between the trigger and the target. For example, we assigned distance=1 for adjacent vowels, 2 for one intervening vowel, and so on. From the identified peak-layer representations, the ATR feature of each target vowel is predicted. We report mean prediction accuracy and classifier confidence at each distance to measure the distance decay. We also computed Pearson's $|r|$ to test linear distance effects, excluding the distance of 5 (n = 30) due to an insufficient sample size. We expect that a model encoding long-distance iterative harmony spreading should show decreasing accuracy with distance. We also examined internal consistency (the frequency of words where all vowels received identical ATR predictions) across 2-vowel, 3-vowel, and 4+-vowel words to assess word-level coherence.

4 Findings

We present the results around two questions: (a) where ATR information is encoded, and (b) whether this encoding reflects grammatical rule learning or statistical co-occurrence.

4.1 Layerwise Probing Analysis

We evaluated ATR encoding across all 24 transformer layers using logistic regression probes. Figure 1 shows classification performance for both models.

For xlsr-53, ATR classification improved steadily through lower and mid layers, peaking at Layer 12 with accuracy of 0.811 ± 0.083, macro-F1 of 0.787 ± 0.081, and ROC-AUC of 0.891 ± 0.072. Performance declined in deeper layers, suggesting that ATR-relevant information becomes less linearly separable as representations encode higher-level structure. xls-r-300m showed a similar but more distributed pattern. Classification accuracy peaked at Layer 16 (0.804 ± 0.079, F1 = 0.781 ± 0.077, AUC = 0.898 ± 0.068), though performance remained high across Layers 11–19. The broader peak suggests that xls-r-300m distributes ATR information across more layers than xlsr-53.

Having established that ATR information is encoded, we tested whether the models learn the grammatical constraints governing harmony. All process tests utilized Layer 12 representations for xlsr-53 and Layer 16 for xls-r-300m, as these layers demonstrated peak accuracy (93.3% for both) in contextual ATR prediction.

4.1.1 Cross-vowel Masking Test: Does the model learn feature agreement?

The cross-vowel masking test assessed whether the model learns that all vowels in a harmonized word share the same ATR feature. The regression probe, trained on the best-performing layers of the models, predicted each vowel's ATR feature

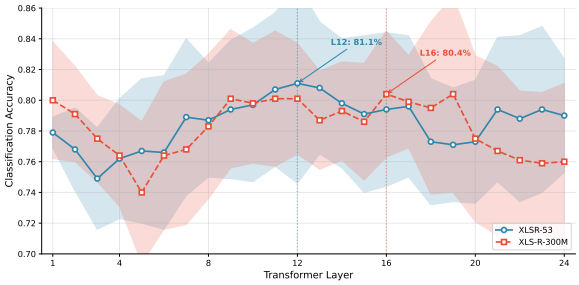

Figure 1: Layerwise ATR classification accuracy for word-level harmony prediction. xlsr-53 peaks at Layer 12; xls-r-300m shows a more distributed pattern a with peak around Layer 16.

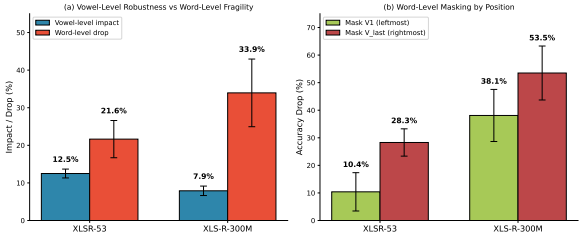

Figure 2: Masking impact at vowel and word levels. (a) Vowel-level encoding is robust ($7.9 - 12.5\%$ affected), while word-level classification is fragile ($21.6 - 33.9\%$ drop). (b) Rightmost vowel masking causes the largest word-level disruption.

relying on contextual information from its neighboring vowels. Across 18100 vowel tokens, the probe achieved 98% mean prediction accuracy for *xlsr-53*($t = 466.75$, p<0.001) and 97.8% for *xls-r-300m*. For word-level agreement and internal consistency, the classifier reached 99.5% (xlsr-53) and 99.3% (xls-r-300m), 98.5% (*xlsr-53*) and 98.2% (*xls-r-300m*) prediction accuracy, respectively (all $p < 0.0001$). Standardized effect sizes range from 3.4 to 6.8 for each of these tests. These results indicate that wav2vec2's intermediate representations encode near-ceiling within-word agreement, an essential property of Assamese vowel harmony.

We further examined how masking affects predictions at the vowel and word levels. At the vowel level, we observed that only 12.5% of other vowels changed predictions for *xlsr-53* and 7.9% for *xls-r-300m*. However, the models differed in positional sensitivity. Masking the rightmost vowel (potential trigger) showed weak directional encoding with significantly larger changes than other positions ($p = 0.0004$) for xlsr-53. In contrast, xls-r-300m showed that the rightmost position had less impact ($p < 0.0001$), suggesting no directional structure.

Word-level masking [2] caused significant accuracy drops in both models (21.6% average for *xlsr-53* and 33.9% for *xls-r-300m*). The rightmost vowel position caused the most significant feature prediction disruption (28.3% and 53.5% accuracy drops, respectively), indicating reliance on the presence of all segments for word-level harmony classification. Our observation suggests that the S3Ms represent ATR feature agreement as a global encoding consistent with statistical learning of surface co-occurrences rather than the acquisition of a harmony rule.

[2]harmonized vs. bare classifier accuracy when one vowel is masked

4.1.2 Directionality Test: Does the model prefer Regressive Harmony?

We designed the directionality test to assess whether the model encodes ATR harmony as a right-to-left (regressive) or left-to-right (progressive) process. We conducted two asymmetric prediction tasks: one test utilized only the right-context information (examining regressive spreading), and the other used only the left-context information (examining progressive spreading).

Both models showed weak but statistically significant regressive bias. *xlsr-53* predicted left vowel's ATR from right context with 72.3% accuracy and right vowel's ATR from left context with 67.7% accuracy, resulting in a directional bias of $+4.7\%$ ($t = 8.94, p < 0.001$). xls-r-300m achieved almost similar results with 72.4% regressive accuracy, 67.5% progressive accuracy, and a bias of $+4.9\%$ ($t = 9.12, p < 0.0001$).

We observed that, although the bias is statistically significant for the respective models, their magnitudes are modest, and the effect sizes are small. Both models show a preference for regressive harmony but do not encode the directional spreading mechanism of Assamese vowel harmony strongly.

4.1.3 Trigger-specificity Test: Do [+ATR] vowels emerge as the trigger?

In Assamese, [+ATR] vowels harmonize [-ATR] vowels, surfacing as the harmony triggers. The trigger specificity test examined whether the model differentiates between [+ATR] and [-ATR] vowels as potential triggers of harmony.

xlsr-53 achieved 89.6% accuracy for words with [+ATR] triggers and 83.6% accuracy for [-ATR] triggers with a difference of $+5.3\%$ (Cohen's $d = 0.24$, small effect). *xls-r-300m* showed a similar

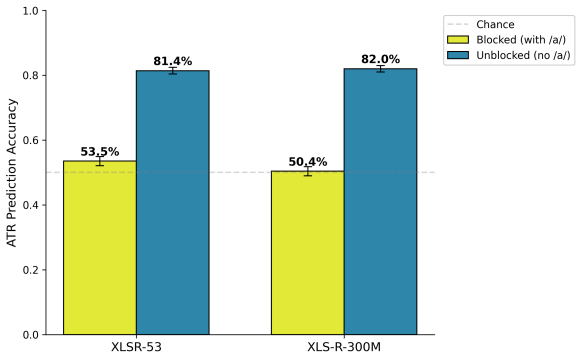

Figure 3: Opacity test results. Both models show significantly lower ATR agreement when /ɑ/ intervenes between vowel pairs, indicating learned blocking. $^{***}p < 0.001$.

prediction accuracy pattern: 90.1% for [+ATR] and 84.0% for [-ATR], yielding a difference of $+6.1\%$ (Cohen's $d = 0.26$).

The trigger-specificity test revealed that the model prefers [+ATR] vowels as triggers, but the impact is weak and subject to variability. This also suggests that trigger specification may emerge from formant trajectories and energy differences, rather than an explicitly learned phonological rule.

4.1.4 Opacity Test: Does the model identify /ɑ/ as a harmony blocker?

We tested the model's ability to recognize /ɑ/ as the opaque vowel (harmony blocker) by comparing harmony prediction accuracy between two contexts: one of them is the blocked condition (where /ɑ/ intervenes between the trigger and target vowels), and the other one is the unblocked condition, where /ɑ/ doesn't intervene.

Both models exhibited strong blocking effects. For *xlsr-53*, vowel pairs separated by /ɑ/ agreed in predicted ATR 53.5% of the time, compared to 81.4% for pairs at comparable distance without /ɑ/. *xls-r-300m* achieved 50.4% prediction accuracy in blocked context and 82.0% in unblocked context. Blocking effect is $+27.9\%(t = 16.2, p < 0.0001$, Cohen's $d = 1.42)$ for *xlsr-53* and $+31.7\%(t = 18.5\%, p < 0.0001$, Cohen's $d = 1.58)$ for *xls-r-300m*. This is calculated as accuracy without a blocker minus accuracy with a blocker. A negative value would indicate that the accuracy is slightly higher when predicting across /ɑ/ than when predicting without /ɑ/ at a comparable distance. The positive value in our case indicates that /ɑ/ reduces harmony accuracy, that is, blocks harmony. Therefore, wav2vec2 representations show lesser ATR agreement when /ɑ/ intervenes between vowels, consistent with sensitivity to opacity.

4.1.5 Distance Decay Analysis: Does the model encode long-distance iterative harmony?

The distance-decay test assessed whether ATR harmony's strength decays as the trigger-target distance increases. We expect that if the model is capable of picking up long-distance iterative patterns, the prediction confidence and accuracy would decrease as the number of intervening vowels increases. We observed that the accuracy remained high and stable across distances 1-4 (xlsr-53: $0.976 - 0.986$; xls-r-300m: $0.970 - 0.980$)[3]. No significant correlation between distance and accuracy was noticed for either model (xlsr-53: r = 0.020, p = 0.17; xls-r-300m: r = -0.005, p = 0.75). The correlation between confidence and distance is also weakly positive (r = 0.066 and r = 0.059, p < 0.001, although this may possibly reflect smaller sample sizes at longer distances.

However, we observed a sharp drop in the assignment of identical ATR values to all vowels across word length: from 74.6% to 61.4% to 41.3% for xlsr-53 (2-vowel, 3-vowel, and 4+-vowel words, respectively), and from 72.9% to 60.4% to 41.1% for xls-r-300m. Individual vowel predictions remain accurate at all distances, but the models struggle to assign consistent ATR values across all vowels in longer words. This pattern, therefore, appears consistent with encoding ATR as a vowel-level property rather than computing iterative spreading from trigger to targets.

5 Conclusion

Our study asked: when a neural speech model accurately distinguishes a phonological contrast, does it actually learn the grammatical rule that governs the contrast? Our findings suggest that the answer is not very straightforward. The understudied wav2vec2.0 representations encode some properties associated with harmony while failing to capture others.

Both the wav2vec2 variants encode ATR distinction in their middle layers. ATR classification peaked at Layer 12 for *xlsr-53* and Layer 16 for *xls-r-300m*. This aligns with previous findings showing

[3]We had 1001 tokens with distance=2 ((C)V1CV2(C),2230 tokens with distance=3 ((C)V1CV2CV3(C), and 1558 tokens with 4+ vowels.

Test	Metric	xlsr-53	xls-r-300m
Layerwise Probing	*Peak layer*	L12	L16
	Peak accuracy	81.1% ± 8.3%	80.4% ± 7.9%
Feature Agreement	*Cross-vowel accuracy*	98.0%	97.8%
	Word-level agreement	99.5%	99.3%
	Internal consistency	98.5%	98.2%
Masking Impact	*Vowel-level* (% affected)	12.5%	7.9%
	Word-level (avg drop)	21.6%	33.9%
Directionality	*Regressive bias*	+4.7%	+4.9%
Trigger Specificity	*[+ATR] advantage*	+5.3% (d=0.24)	+6.1% (d=0.26)
Opacity	*Blocking effect*	+27.9% (d=1.42)	+31.7% (d=1.58)
Distance Decay	*Distance-accuracy correlation*	r=0.020 (p=0.17)	r=−0.005 (p=0.75)
	Internal consistency (4+ vowels)	41.3%	41.1%

Table 2: Summary of layerwise probing and process-level test results for both models.

that phonological information is encoded in the intermediate layers following acoustic process but preceding higher-level contextual encoding (Pasad et al., 2021). Probing results alone might suggest successful learning of ATR vowel harmony. However, when representations are evaluated using masking-based diagnostics that target process-level properties, the picture becomes more complicated.

Assamese ATR harmony, like any harmony system of the world, is not just based on within-vowel agreement. It is a process in which [+ATR] vowels spread their feature in the right-to-left direction, with /ɑ/ blocking the spread. A model that has truly learned harmony must encode these phonological properties. It should reflect the regressive bias, treat [+ATR] vowels as the trigger, show iterative long-distance harmony, and learn /ɑ/ as the opaque vowel.

Both models strongly encode within-word ATR feature agreement and show clear sensitivity to the presence of an intervening low vowel /ɑ/. In contrast, evidence for directional spreading, trigger asymmetry, and iterative computation is weak or absent. The process-level results reveal that the models predict more accurately in a regressive direction than in a progressive one (directional bias $+4.7\%$ to $+4.9\%$), and [+ATR] trigger scores modestly higher accuracy than [-ATR] contexts (Cohen's $d = 0.24 - 0.26$); however, none of their effect sizes are significantly large. The opacity findings reflect a large drop in prediction accuracy when /ɑ/ intervenes between vowels, with blocking effects of $28 - 31\%$ with medium to large effect sizes (Cohen's $d = 1.42 - 1.58$). However, we cannot rule out that this reflects acoustic dissimilarity of /ɑ/ in embedding space rather than learned phonological opacity. Theoretically, blocking is a consequence of directional spreading being interrupted. In our case, a strong opacity effect along with a weak directionality effect indicates that the models capture local differences without representing the directional process. These dissociative results suggest that hi h probing accuracy is not equivalent to grammatical knowledge.

Wav2vec2, as a transformer model, applies a self attention over the entire input win ow without an inherent bias toward sequential processing (Vaswani et al., 2017; Baevski et al., 2020). This global mechanism favors encoding the feature co-occurrence patterns over directional dependencies. Furthermore, the pretraining objective of contrastive loss encourages discriminative representations rather than rule-inductive ones (Baevski et al., 2020). This may also explain why /ɑ/ emerges robustly as a blocker, but directional spreading, which requires rule-informed sequences, is weakly represented. Future work could test this hypothesis by examining correlations between acoustic measures (F1, F2, spectral properties) and model predictions, or by probing whether models trained on languages with progressive harmony show reverse directional biases. Targeted masking strategies such as comparing predictions in cases of partial spreading or manipulating distance and intervening context in blocked versus unblocked words could further reveal whether masking-induced disruption reflects learned grammatical constraints or statistical distributional biases.

Limitations

We acknowledge the limitations in our work. We have worked on a single language with approxi-

mately 2 hours of speech from 14 adult speakers. Additionally, our dataset consists of isolated words. It would be worthwhile to investigate how the models behave on continuous speech and whether our observations can be generalized to other languages that exhibit vowel harmony. Also, we compared two multilingual variants of wav2vec2. Extending this analysis to other state-of-the-art architectures would strengthen the generalizability of our observations.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*, pages 2278–2282.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic. *arXiv preprint arXiv:2406.08619*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*.

Domenico De Cristofaro, Vincenzo Norman Vitale, and Alessandro Vietti. 2025. Evaluating the representation of vowels in wav2vec feature extractor: A layer-wise analysis using mfccs. *Preprint*, arXiv:2508.17914.

Patrick Cormac English, John Kelleher, and Julie Carson-Berndsen. 2022. Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91.

Patrick Cormac English, John D Kelleher, and Julie Carson-Berndsen. 2023. Discovering phonetic feature event patterns in transformer embeddings. In *Proc. Interspeech*, pages 4733–4737.

Patrick Cormac English, John D Kelleher, and Julie Carson-Berndsen. 2024. Searching for structure: Appraising the organisation of speech features in wav2vec 2.0 embeddings. In *Interspeech*, volume 2024, pages 4613–4617.

Jon Gauthier, Canaan Breiss, Matthew K Leonard, and Edward F. Chang. 2025. Emergent morpho-phonological representations in self-supervised speech models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28055–28074, Suzhou, China. Association for Computational Linguistics.

Sai Gopinath and Joselyn Rodriguez. 2024. Probing self-attention in self-supervised speech models for cross-linguistic differences. *Preprint*, arXiv:2409.03115.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Ashish Jaiswal, Ashwin ramesh babu, Mohammad Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9:2.

Shakuntala Mahanta. 2007. *Directionality and locality in vowel harmony*. PhD Thesis, Utrecht University.

Shakuntala Mahanta. 2008. *Directionality and locality in vowel harmony: With special reference to vowel harmony in Assamese*. Netherlands Graduate School of Linguistics.

Shakuntala Mahanta. 2012. Assamese. *Journal of the International Phonetic Association*, 42(2):217–224. Publisher: Cambridge University Press.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. In *Proc. Interspeech 2023*.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.