To agree or not to agree? Evaluating measures of similarity in consonant harmony*

Lydia Quevedo & Kate Mooney
University of Maryland, College Park

1 Introduction

In languages with consonant harmony, consonants which are highly similar in articulatory, perceptual, or featural properties are required to agree by the value of some feature (Gafos, 1996; Hansson, 2010; Mackenzie, 2011; Rose and Walker, 2004). Previous work has proposed similarity metrics to describe these classes of highly similar consonants (Pierrehumbert, 1993; Frisch et al., 2004). However, these metrics have yet to be applied predictively to determine which consonants should participate within a particular language's harmony system. In this paper, we evaluate different similarity metrics to explain why certain segments participate in harmony within a given language.

1.1 *An illustration of the problem.* For a concrete example, consonant harmony in Shilluk (Nilotic, Gilley 1992) requires that coronal stops and nasals match in [±ANTERIOR], as shown in (1). A dental stop or nasal can only be followed by a dental stop or nasal, likewise for alveolar stops and nasals.

(1) Shilluk: Roots contain only alveolar or dental stops (Gilley, 1992)

	Dentals			*Alveolars*	
a.	àd̪út̪	'stinger'	b.	dút	'loin cloth'
c.	t̪in̪	'small'	d.	tin	'today'

Only coronal stops and nasals participate in harmony, and only these same segments contrast for dentality, as indicated by the shading in (2).

(2) Shilluk consonant inventory (Gilley, 1992:28)

		Labial	Dental	Alveolar	Palatal	Velar
Plosives	*Voiceless*	p	t̪	t	c	k
	Voiced	b	d̪	d	ɟ	g
Nasals		m	n̪	n	ɲ	ŋ
Liquids		w		l	j	
Vibrant				r		

The question then is how the contrasts of the inventory are connected to the kind of harmony observed in a particular language. For instance, there are other consonants in Shilluk that could plausibly undergo harmony, such as the coronal liquids. In other languages like Hausa (Newman, 2000), /l, n, r/ participate in lateral harmony, so that when an /l/ is followed by another coronal sonorant, that consonant must also be /l/; see details in (3) and the full inventory in (4):

(3) Hausa: Harmony in sonorants and homorganic obstruents (Newman, 2000; Hansson, 2010)

Type	*Description*	*Participants*
Sonorant	/l/ cannot co-occur with /n, ɽ/	/l, n/, /l, ɽ/
Sonorant	*/n...ɽ/ but /ɽ...n/	/ɽ, n/
Laryngeal	Homorganic cons. must agree in phonation	/b, b, d, d, s, s', k, k'/

[*] Many thanks to Naomi Feldman, William Idsardi, Joselyn Rodriguez, members of the UMD Linguistics Department, and members of the AMP 2025 audience for their helpful comments.

Proceedings of *AMP 2025*

(4) Hausa consonant inventory (Newman, 2000:392)

		Labial	Alveolar	Palatal	Velar	Glottal
Plosives	*Voiceless*	p	t^h	c^h	k^h	ʔ
	Voiced	b	d	ɟ	g	
	Glottalic	b	d		ƙ	
Nasals		m	n			
Fricatives	*Voiceless*		s	ʃ		h
	Voiced		z			
	Glottalic		s'			
Affricates				tʃ dʒ		
Tap			ɽ			
Glides		w			j	
Lateral			l			

Despite also having these coronal sonorant contrasts, Shilluk lacks lateral harmony, and so we can conclude that the presence of a contrast is not enough to predict which consonants participate in harmony in a given language.

1.2 *Previous work: Harmony and similarity.* Previous work has claimed that consonants that participate in harmony are always similar. Rose and Walker (2004) argue that similarity is evaluated over features, and crucially, that not all features are weighed equally. Languages are more likely to rank CORR-CC high for segments that match in [SON], [CONT], PLACE (after Pierrehumbert (1993)).

(5) *Agreement by correspondence* (ABC) (Rose and Walker, 2004)

a. CORR-CC: consonants above a certain threshold of similarity must correspond. *(our wording)*

b. IDENT-CC[F]: corresponding consonants must agree by the value of [F]

Outside of harmony, previous work has also claimed that segments have co-occurrence restrictions when they are featurally similar. For instance, Frisch et al. (2004) examine OCP-Place effects in Arabic, where two phonemes cannot co-occur within the same root. The more natural classes the two phonemes share, the stronger the co-occurrence restriction between them.

Doucette et al. (2024) also examine consonant co-occurrence restrictions within 107 languages (21 families) of the NorthEuraLex database. They also observe that similarity, as computed over shared natural classes, also predicts consonant co-occurrence restrictions. As similarity increases, so does the strength of the co-occurrence restriction. These effects were found gradiently in the lexicon of all languages surveyed, even though only one language, Basque, had consonant harmony as a categorical effect.

Beyond Rose and Walker (2004), the link between consonant harmony and similarity has proven problematic. Mackenzie (2011, 2016) observes that natural class similarity (as developed by Frisch et al. 2004) does not always predict which consonants harmonize. In Nilotic, for example, both Anywa and Dholuo have very similar consonant inventories, but differ in whether /n/ participates in dental harmony. Mackenzie proposes that segments only participate in harmony if they are specified for the relevant feature, and that the difference between Anywa and Dholuo is in whether /n/ is specified for [ANT] or underspecified. Only specified consonants participate. While this analysis successfully describes the Nilotic facts, there is no evidence independent of harmony that motivates underspecification in these languages, and so the analysis is not predictive.

Arsenault (2012) conducts a thorough typology of harmony and dissimilation in retroflex consonants of South Asia. Many of these languages have co-occurrence restrictions on retroflexes, so that retroflex consonants do not combine freely with other coronals. Arsenault observes that neither natural class similarity nor Mackenzie's underspecification can correctly predict which consonants participate in these co-occurrence restrictions. For instance, both Kalasha and Indus Kohistani have retroflex harmony between their obstruents. But, Kalasha permits co-occurrence of palatal affricates and retroflex sibilants

(✓[c... s]), but Indus Kohistani does not (✗[c... s]). Existing similarity metrics incorrectly predict that these languages should have the exact same harmony pattern, because their inventories are quite similar.

We are left with a problem of implementation. Similarity seems to capture the broad typological generalizations about where consonant harmony occurs. However, existing similarity measures fail to correctly predict which consonants participate in harmony.

1.3 *Hypotheses.* To date, there is no predictive account of which consonants in a given language participate in harmony. The aim of this paper is to establish a baseline over possibilities developed in the literature, where we determine if existing measures of similarity can predict which consonants harmonize.

Based on the literature, we present two main hypotheses:

(6) Strong claim: Consonants which participate in harmony are more similar than any other pair.

a. **Global similarity hypothesis:** If consonant pairs differing in a dimension α are more similar than all other pairs differing in any dimension β, then harmony applies over α.

b. **Within-class similarity hypothesis:** When grouped by natural classes, if pairs differing in a dimension α are more similar than all pairs differing in any dimension β, then harmony applies over α.

We test these hypotheses under three main definitions of similarity: **(1) Phonological features**, using (a) universal features from Panphon (Mortensen et al., 2016), (b) underspecified features starting from stricture, and (c) underspecified features starting from major Place; **(2) Surface representations (with SSL speech models)**, and **(3) A combination of surface representations and phonological features**.

We illustrate these approaches with two case studies, (1) dental harmony in Shilluk and (2) laryngeal/lateral harmony in Hausa (see Section 1.1). We find that no similarity measure perfectly predicts the class of harmonizing consonants for both languages. While featural similarity is predictive of consonant harmony patterns in Shilluk, surface representations are more predictive for Hausa. For the languages examined, there is no similarity measure which universally captures attested consonant harmony patterns.

1.4 *Roadmap.* The paper is organized as follows: Section 2 on similarity over phonological features, using both a universal fully-specified feature matrix from PanPhon and two underspecified feature matrices. Section 3 evaluates similarity over surface representations. Section 4 combines these two approaches, using surface similarity over natural classes of consonants (e.g. homorganic obstruents). Section 6 concludes.

2 Study 1: Similarity over features

2.1 *Method.* The question at hand is whether similarity computed over phonological features is predictive of participation in consonant harmony. Our approach closely follows Doucette et al. (2024), where similarity is computed over phonological features using two different algorithms. we call these *feature counting* (Pierrehumbert, 1993) and *natural classes* (Frisch et al., 2004).

(7) Feature counting similarity metric

$CountSim(x,y) = \frac{|\mathrm{F(x)} \bigcap \mathrm{F(y)}|}{|\mathrm{F(x)} \bigcup \mathrm{F(y)}|}$

(Pierrehumbert, 1993)

(8) Natural class similarity metric

$ClassSim(x,y) = \frac{|\mathrm{NC(x)} \bigcap \mathrm{NC(y)}|}{|\mathrm{NC(x)} \bigcup \mathrm{NC(y)}|}$

(Frisch et al., 2004)

We evaluate these similarity measures over three different feature systems. First, we use Panphon (Mortensen et al., 2016), a universal feature system that fully specifies the feature values for every phoneme in a language's inventory. We also add two underspecified feature matrices, one which subdivides by place first, and another that subdivides by stricture first. The two underspecified feature matrices are generated using an R script R Core Team (2025) following a modified version of the Successive Division Algorithm (SDA) (Dresher, 2003, 2009). The basic procedure is described in (9).

(9) Procedure for generating underspecified feature matrices

1. Begin with all phonemes in the inventory.

2. Starting with the first feature [F], assign values for [F] and divide the inventory into N groups according to each value of [F]. For each group G:
   (a) See if there is more than one phoneme in G. If yes, proceed to (2b). If no, proceed to (2d).
   (b) Scan the feature list (*see (10)-(11*) for the next feature [F'] which is contrastive for at least two members of G. Assign a value for [F'] to each member of G.
   (c) Divide G into N subgroups for each value for [F']. For each subgroup, return to (2a).
   (d) For all remaining features, assign the value 0.

Features are assigned using one of two fixed orders: *Place-first* (following Frisch et al. 2004) and *Stricture-first* (following Rose and Walker 2004). In each of these approaches, either Place or continuancy is assigned first, with subsequent features receiving maximally underspecified feature values. This method is intended to incorporate previous proposals' claims that similarity only relevant within place of articulation / stricture class.

(10) *Place first* Based on Frisch et al (2004)
[syl] ≫ **Lab** ≫ **Cor** ≫ **Dor** ≫
[son] ≫ **[cont]** ≫ [nas] ≫ [lat] ≫ [ant] ≫ [dist] ≫ [voi] ≫ [stri] ≫ [s.g] ≫ [c.g.]

(11) *Stricture first* Based on Rose & Walker (2004)
[syl] ≫ **[son]** ≫ **[cont]** ≫
Lab ≫ Cor ≫ Dor ≫ [nas] ≫ [lat] ≫ [ant] ≫ [dist] ≫ [voi] ≫[stri] ≫ [s.g] ≫ [c.g.]

For illustration, the Panphon (fully specified) feature matrix for a subset of Shilluk consonants is shown in Table 1. Compare with the the matrix in Table 2, which first groups by major Place specification, and only has specified values for features which are contrastive within each place of articulation. In Shilluk, these two matrices differ only slightly, such as in the [cont] value for [ŋ].

Table 1: Panphon feature matrix for a subset of Shilluk consonants

	[Place]	[son]	[cont]	[nas]	[dist]	[voi]
/p/	Lab	-	-	-	-	-
/m/	Lab	+	-	+	-	+
/t/	Cor	-	-	-	-	-
/d/	Cor	-	-	-	-	+
/n/	Cor	+	-	+	-	+
/t̪/	Cor	-	-	-	+	-
/d̪/	Cor	-	-	-	+	+
/n̪/	Cor	+	-	+	+	+
/l/	Cor	+	+	-	-	+
/ŋ/	Dor	+	+	-	-	+

Table 2: Contrastively underspecified by place (left) and stricture (right) for Shilluk.

	[Place]	[son]	[cont]	[nas]	[dist]	[voi]
/p/	Lab	-	0	0	0	-
/b/	Lab	-	0	0	0	+
/m/	Lab	+	-	0	0	0
/w/	Lab	+	+	0	0	0
/t/	Cor	-	0	0	-	-
/d/	Cor	-	0	0	-	+
/n/	Cor	+	-	0	-	0
/t̪/	Cor	-	0	0	+	-
/d̪/	Cor	-	0	0	+	+
/n̪/	Cor	+	-	0	+	0
/l/	Cor	+	+	0	0	0
/ŋ/	Dor	+	0	0	0	0

	[son]	[cont]	[Place]	[nas]	[dist]	[voi]
/p/	-	0	Lab	0	0	-
/b/	-	0	Lab	0	0	+
/m/	+	-	Lab	0	0	0
/w/	+	+	Lab	0	0	0
/t/	-	0	Cor	0	-	-
/d/	-	0	Cor	0	-	+
/n/	+	-	Cor	0	-	0
/t̪/	-	0	Cor	0	+	-
/d̪/	-	0	Cor	0	+	+
/n̪/	+	-	Cor	0	+	0
/l/	+	+	Cor	0	0	0
/ŋ/	+	-	Dor	0	0	0

Over these three feature matrices, we compute similarity with the Count-Sim and Class-Sim metrics described earlier in this section. In the actual computation of similarity, pairs that differ in specified values of features are more similar than pairs that differ by a contrastive value and an underspecified value. Put more formally, a pair contrasting in [+F] and [−F] is more similar than a pair that is [±F] and [0F].

In order to interpret these similarity measures, we need to compare them against some baseline. We use a logistic regression model randomly assigning phonemes as participants or non-participants in consonant harmony patterns. If the featural similarity measures are more predictive than one or both of these baselines, then we may conclude that consonants which participate in harmony patterns are indeed more similar than random.

Lastly, if we want to evaluate how well a given similarity measure predicts participation in harmony, we must precisely define how we intend to make the comparison. We define *participating pairs* as pairs of consonants which have a co-occurrence restriction against them. For example, in Shilluk, six consonants are subject to a co-occurrence restriction on dentality: /t, d, n, t̪ d̪, n̪/. Only some pairs of consonants in this set have co-occurrence restrictions against them. Thus, /t-d̪/ is a participating pair because harmony must be enforced for this pair; it would be a violation of the co-occurrence restriction. Meanwhile, /t-d/ is a *non-participating* pair because they are not subject to any co-occurrence restriction. The present study aims to evaluate whether participating pairs are more similar in Count-Sim or Class-Sim measures than any other pairs of consonants in the inventory.

2.2 *Results: Global Similarity.* We start with the results for Shilluk. Neither similarity score categorically predicts participation in consonant harmony, regardless of the feature matrix used, although there is some variation as to the specifics. There is no difference between underspecified feature matrices, but both underspecified matrices differed from Panphon in the same respects. When similarity is computed over the underspecified matrices, certain non-participating pairs which contrast in voicing are more similar than participating pairs, e.g. /t-d/ is more similar than /n̪-d/ with the Class-Sim measure. Computing similarity over Panphon features results in similar issues. For example, /ɳ̣-ŋ/ and /n̪-n/ both have a Count-Sim score of 0.8333.

The story is much the same for Hausa: There are a number of non-participants which are more similar than participants, particularly heterorganic pairs, e.g. /ɓ-ɗ/ > /n-l/. This is true of all three feature matrices. With Panphon in particular, though, non-participating /j-l/ > participating /ɽ-n/. In conclusion, all versions of the Global Similarity Hypothesis are false, and similarity in phonological features is not absolutely predictive of participation in consonant harmony for Shilluk and Hausa.

At this point, we must ask about the broader typology. Perhaps the Global Similarity Hypothesis is false for Shilluk and Hausa, but true more generally of languages with consonant harmony. To test this hypothesis, we examined eight additional languages with consonant harmony (Anywa, Dholuo, Kikongo, Lezgian, Sawai, Tiene, and Yucatec). In none of the eight languages is similarity globally predictive of participation in consonant harmony. For each case, there is at least one non-participating pair which is more similar than a participating pair. Overall, then, the Global Similarity Hypothesis is not true, because for all languages tested, there are disharmonic pairs (i.e. pairs from the disharmonic class) which are more similar than harmonic pairs.

2.3 *Results: Rank Similarity.* To test how well each similarity measure (COUNT-SIM and CLASS-SIM), predicts participation in harmony, we fit a logistic regression model to the similarity measures. Then, using the resulting probability scores, we conducted a Wilcoxon or Mann-Whitney U-test in Python using the `scipy` package to compute the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values for each model by consonant pair subgrouping. The AUC represents the probability that the true positives or targets (i.e. participating pairs) are ranked above false positives or distractors (non-participants) when sorted by the similarity measure. Thus, an AUC of 1.00 means 100% of the true positives were categorized as participants, with zero false positives, whereas an AUC of 50% means the model performs at chance. The AUCs are reported Table 3. AUC scores with $p < 0.05$ are marked with one asterisk (*), $p < 0.01$ with two (**), $p < 0.001$ with three (***), and insignificant values with gray shading. The ROC curves, which plot the false positive rate are shown in Figure 1.

When sorted by Count-Sim and Class-Sim scores, participating Shilluk consonant pairs are ranked above (i.e. as more similar than) non-participating pairs significantly more often than by chance (AUC

= 0.5) for all consonant pairs and all homorganic pairs. Although the AUCs for both measures are above chance with the two underspecified feature matrices, only Count-Sim is above chance for the Panphon matrix. No combination of similarity measure and feature matrix has a significant AUC when looking at consonant pairs which share sonorancy and/or continuancy, however.

Table 3: Area-under-the-curve (AUC) scores for each measure by language

	Shilluk			Hausa		
	Panphon	Place-first	Stricture-first	Panphon	Place-first	Stricture-first
Count-Sim	0.7004*	0.8554***	0.8195***	0.6927	0.8912***	0.8912***
Class-Sim	0.6808	0.8797***	0.8128***	0.6859	0.8929***	0.8929***

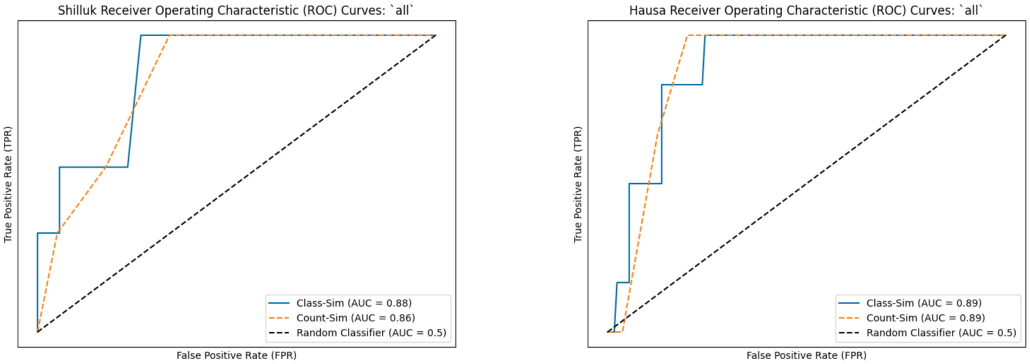

Figure 1: ROC curves for the featural similarity measures for all Shilluk (left) and Hausa (right) consonant pairs.

Turning to Hausa, here we see that the AUCs for neither similarity measure computed over the Panphon matrix are significantly above chance. When computed over the underspecified feature matrices, however, the AUCs are significantly above chance. Thus, similarity is only predictive of participation of harmony when computed over contrastively underspecified features.

Between the two similarity measures, the AUCs differ depending on the feature matrix used. To determine whether these differences are significant, we computed DeLong's (DeLong et al., 1988) AUC test using the `pROC` package (Robin et al., 2011) in RStudio (R Core Team, 2025) to compare these two featural similarity measures to each other and the acoustic baseline discussed above. The results show that the difference in AUCs is not significant in any condition ($p > 0.05$). There was no significant difference in AUCs for any of the feature matrices for Shilluk. We can then conclude that while the Count-Sim and Class-Sim measures are both predictive of participation in consonant harmony, the results are the same for the PanPhon and underspecified feature matrices. Meanwhile, for Hausa, the AUCs for both underspecified matrices were significant, but Panphon was not. DeLong's test reveals that the AUCs for both underspecified matrices were not significantly different. Furthermore, it also shows no difference in AUCs between Class-Sim and Count-Sim measures.

In sum, while featural similarity is not globally predictive of participation in consonant harmony, featural similarity is a significantly better predictor of participation than chance in both languages. The underspecified feature matrices fared better (i.e. had significant AUCs for more subgroupings of consonant pairs) although there was no difference in AUCs between the two.

3 Study 2: Similarity over surface representations

3.1 *Method.* In this study, we compute similarity over low-level representations of the acoustic signal generated by self-supervised (SSL) speech models. The representations generated by SSL speech models have a number of properties that make them a useful tool for phonological research. They have been found to correlate closely with acoustic properties (Pasad et al., 2023) as well as EMA traces (Cho et al., 2023). In turn, these properties end up being remarkably human-like. For example, they display a certain degree of language-specificity in L2 contexts, as a human learner (Rodriguez et al., 2024). Furthermore, they have been found to encode phonetic and phonemic cues for contrasts used by humans (Martin et al., 2023). Even perceptual space generated from these representations has been found to reflect human similarity judgments (Chernyak et al., 2024).

For this study, we use three SSL models which differ in the nature of their pre-training. All three versions use transformer architecture, i.e. they are encoder-decoder models. Two are models which have been pretrained on audio data: a monolingual SSL model (HuBERT) trained on English, and the second a multilingual SSL model (XLS-R) trained on more than a hundred other languages. As a baseline, following Martin et al. (2023), the third is the base version of the multilingual model (Wav2Vec2) initialized with random weights. All pretrained versions of the models were accessed using the Hugging Face `transformers` Python library. The attention weights were not fine-tuned to the language data being used here. Additional technical details about these models are shown in (4). These models all take raw audio files as their input and output vector representations (embeddings) of the audio.

The basic procedure is as follows. First, raw audio files are resampled to 16k Hz. Second, embeddings are extracted for the audio files from all transformer (i.e. hidden) layers of the models. Then, embeddings are averaged together within phonemes. Finally, the average pairwise distance is computed between phoneme categories.

Table 5 provides an overview of the sources for the audio and TextGrid corpora used for this study. Note that the audio data for Hausa comes from Common Voice, an audio corpus which was used in the pretraining process for the multilingual model, XLS-R. Thus, if there is a difference in acoustic similarity measures from XLS-R for these two languages, then this may be at least partially attributed to the fact that it has been trained on data from Hausa but not Shilluk. Specifically, having been trained on specific types of contrasts is makes a major difference in how these models generate representations of acoustics, then we would expect the similarity measures for Shilluk to be worse at predicting participation in harmony than for Hausa, especially in the case of the monolingual model. If the multilingual model does indeed generate representations which generalize across languages, even when those languages did not constitute a part of the training dataset, then it should perform comparably with both languages.

Table 4: Technical details about the SSL speech models

	Monolingual model	Multilingual model	Baseline
Name	HuBERT (Hsu et al., 2021)	XLS-R (Babu et al., 2021)	Wav2Vec2 (Baevski et al., 2020)
Checkpoint	`hubert-large-ll60k`	`wav2vec2-xls-r-2b`	`wav2vec2-base-960h`
Pre-training	60k hours of English Generates labels using k-means, then guesses for masked portion of audio.	436k hours (128 langs) Predicts label for masked portion of audio.	None

Table 5: Audio and TextGrid corpus details

	Shilluk	Hausa
Harmony	Dental	Laryngeal, liquid
Audio source	Gwado Ayoker and Remijsen (2017) (13 minutes)	Common Voice 16 2020 (11 minutes)
TextGrid source	Hand alignment	Forced alignment (Ahn & Chodroff 2022)

3.2 *Results: Global Similarity Hypothesis.* In this section, we report the findings for similarity in cosine distance, i.e. Jaccard similarity. Since the earliest transformer layers have been found to correlate most closely with acoustic properties of speech, i.e. MFCCs (Pasad et al., 2023; Martin et al., 2023), we will examine the results for layer `i=2` in both the monolingual and multilingual models.

For both the monolingual and multilingual models, the Global Similarity hypothesis is false. That is, there are pairs of consonants which do not participate in harmony and yet are more similar in their acoustic properties than pairs that do in fact participate in harmony. There are fewer problematic pairs like this in Hausa than in Shilluk, and the pairs in Hausa can be dealt with if we make certain assumptions.

In Shilluk, some pairs that contrast in voicing are more similar than certain pairs that contrast in dentality, i.e. participating pairs. For example, the alveolar stops /t-d/ and palatal stops /c-ɟ/ are more similar than /t̪-t/ and /n̪-d/, both pairs that participate in harmony. Pairs like /p-t/ and /b-g/ which contrast in place of articulation are also more similar than all pairs that participate in harmony. Similarly, in Hausa some pairs contrasting in major place, like /ɓ-d/, are also more similar than homorganic pairs like /b-ɓ/ that contrast in laryngeal specifications and are thus subject to the harmony patterns in (3). We can thus conclude that the Global Similarity Hypothesis is false for both languages.

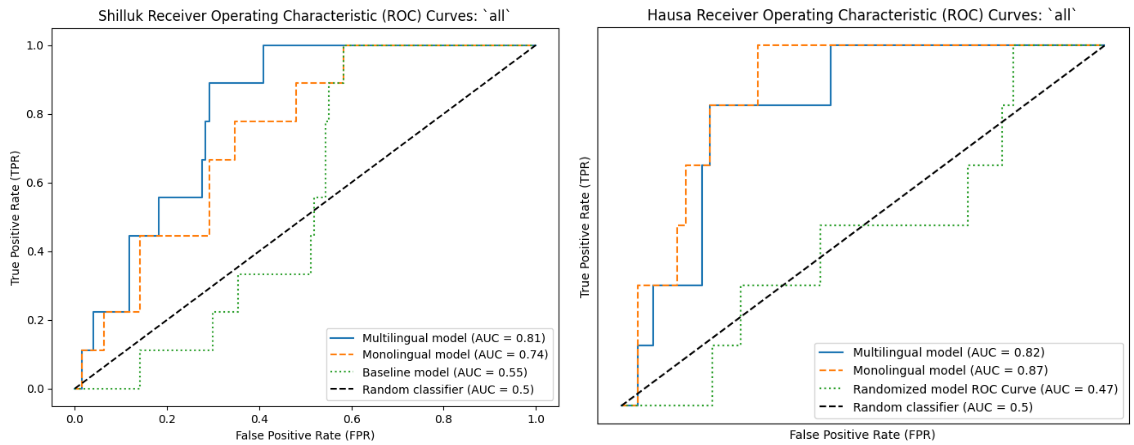

Figure 2: ROC curves for cosine similarity of SSL representations for Shilluk (left) and Hausa (right) with all consonant pairs

3.3 *Results: Rank Similarity Hypothesis.* We can now consider the possibility that while the pairs of consonants that participate in harmony may not be universally more similar in acoustic properties than any other pairs of consonants, the participating pairs may have a higher probability as a class of being ranked higher in similarity than the pairs which do not participate in harmony.

As in the previous study, we conducted a Wilcoxon or Mann-Whitney U-test in R on the resulting probability scores to compute the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) for each model. The resulting area-under-the-curve (AUC) scores are reported in Table 8, while the ROC curves are shown in Figure 2.

Table 6: Area-under-the-curve (AUC) scores for each model by language

Shilluk			Hausa		
Monolingual	*Multilingual*	*Baseline*	*Monolingual*	*Multilingual*	*Baseline*
0.7384**	0.8075**	0.5503	0.8694**	0.8250***	0.4725

As we can see, the baseline model is never above chance (AUC = 0.5) in any condition for either language. Both the monolingual and multilingual models are significantly above chance. That is, surface similarity is significantly predictive of participation in consonant harmony.

For each language, the AUC scores differ between the two pretrained SSL models. To test whether these are significant, we performed DeLong's test, as in Study 1 (Section 2). With Shilluk, the multilingual model is significantly better (p = 0.009) than the monolingual one. Meanwhile, the reverse is true for Hausa: the monolingual model is better (p = 0.002). Although similarity in surface properties is predictive of participation in consonant harmony, no one model works for both languages.

3.4 *Discussion.* There are many possible reasons why SSL similarity measures are predictive of consonant harmony patterns in Hausa but not Shilluk. If pretraining experience with a particular language is important, this may be one source of the difference: The multilingual model was trained on a dataset that included the same Hausa audio corpus used in Study 2, but no Shilluk audio, nor data from any other Nilotic language (Babu et al., 2021). Meanwhile, the pretraining corpus for the monolingual model only includes data from English (Hsu et al., 2021). We would then expect that 1) the multilingual model should generalize better Hausa than Shilluk, and 2) the multilingual model should perform better on Hausa than the monolingual one. We saw in Study 2 (Section 3) that the multilingual model was better than the monolingual model for Shilluk, but there was no difference between the models for Hausa. This means that pretraining on a specific language may not be as important for representing the surface properties relevant to consonant harmony.

Alternatively, perhaps what matters is not experience with a specific language, but rather exposure during pretraining to a particular type of contrast. If the pretraining dataset contains more examples of manner contrasts among Coronal sonorants (e.g. /l, r, n/) than contrasts involving [±DISTRIBUTED] or another fine distinction in tongue tip contact location/orientation (e.g. /t-t/, /t-ʈ/), then perhaps the embeddings more faithfully encode surface properties of the sonorant contrasts than the tongue tip contrasts. Still, this is not a satisfactory explanation, because the multilingual model was pretrained on several languages with [DISTRIBUTED] contrasts (Malayalam, Tagalog, Tamil, Telugu, Thok Reel...), meaning that it is far from inexperienced with these contrasts. An interesting avenue for future research would be examining what specific aspects of pretraining influence a model's representations of surface properties. For now, it is clear that the presence or absence of certain information alone does not explain the difference between Shilluk and Hausa in this study.

4 Study 3: Surface representations and features

4.1 *Method.* In the previous study, pairs of participating consonants were significantly more similar in surface properties as computed with the monolingual SSL model than non-participating pairs. But perhaps the model is able to trivially distinguish the two classes based on factors other than their surface properties. As mentioned above, the six consonants subject to Shilluk co-occurrence restrictions are /t, d, n, t̪ d̪, n̪/. They differ from each other by only a few features: [SONORANT], [NASAL], [DISTRIBUTED], and [VOICE]. They also share a feature: [-CONTINUANT], one of the features that Rose and Walker highlight as an important to consonant harmony. The remaining thirteen consonants in Shilluk are /p, b, m, w, l, r, c ɟ, ɲ, j, k, g, ŋ/ which are also non-participants. These consonants differ from each other in [PLACE], [CONSONANTAL], [CONTINUANT], [SONORANT], [ROUND], and [VOICE]. As they do not share sonorancy or continuancy, they are not sufficiently similar in Rose and Walker's view to be subject to harmony. They are also quite distinct in their acoustic properties. Given these significant differences in featural and phonetic properties, compared to the set of participating consonants, then any similarity measure may trivially be able to distinguish between participants and non-participants. A fairer comparison needs to control the number and type of feature differences between pairs. Our third study aims to address this issue.

Table 7: Summary of pair subgroupings

	Share...	Differ in...
`all`	Any feature(s)	Any feature(s)
`hom`	PLACE	Manner, [CONT], [SON], laryngeal
`hom-either`	[PLACE] AND ([CONT] OR [SON])	Manner, [CONT], [SON], laryngeal
`hom-both`	[PLACE], [CONT] AND [SON]	Manner, laryngeal

We created four groups of consonant pairs: `all`, `hom`, `hom-either`, and `hom-both`. Each group was designed to incorporate various assumptions from the literature. Firstly, `all` is the control group, composed of all consonant pairs in the language. Next, HOM includes only homorganic consonant pairs, following Frisch et al. (2004). As previously mentioned, one of Rose and Walker's core observations is that the members of the harmonic class usually share either [SON] or [CONT] or both. Since it is not clear to what extent the harmonic class must share both, there is a consonant pair subgrouping to cover each possibility. `hom-either` includes the homorganic pairs which share either [SON] or [CONT] or both, while `hom-both` is the set of homorganic pairs which share both of these features. Table 7 summarizes which features must match and which features may differ for each subgrouping.

4.2 *Results.* Are participating consonants are similar in their surface properties out of the set of consonants which are comparable in homorganicity and/or sonorancy/continuancy properties? We focus on the monolingual and multilingual models as they were both above chance in Study 2. Below, Table 8 shows the AUC scores from the the Mann-Whitney test, and Figure 3 shows the ROC curves for the consonant pair subgrouping `hom-both`, all pairs matching in [SON] and/or [CONT].

Table 8: Area-under-the-curve (AUC) scores for each subgrouping by language

	Shilluk		Hausa	
	Monolingual	*Multilingual*	*Monolingual*	*Multilingual*
`all`	0.7384**	0.8075**	0.8694**	0.8250***
`hom`	0.6889	0.7689***	0.9333**	0.9000***
`hom-either`	0.6257	0.7047	0.9636**	0.9273***
`hom-both`	0.6933	0.7467	1.00**	1.00**

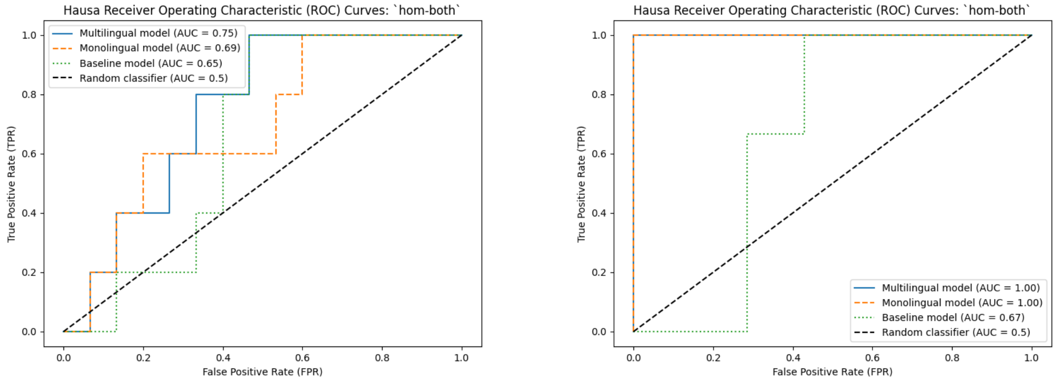

Figure 3: ROC curves for cosine similarity of SSL representations for Shilluk (left) and Hausa (right) homorganic consonant pairs matching in both [SON] and [CONT]

Starting with Shilluk homorganic pairs, only the multilingual model (AUC = 0.77; p = 0.009) is significantly better than chance; the monolingual model is not significant (AUC = 0.69; p > 0.05). None of the models are significant with homorganic consonants which share [SON] or [CONT]. Only the similarity scores from the multilingual model are predictive of participation of consonant harmony.

For the most part, the AUCs for the multilingual model are greater than or equal to the AUCs for the monolingual model in each homorganic condition. DeLong's test shows that for Shilluk, the AUC for the multilingual model is significantly higher than the monolingual model for subgroupings `hom` (p < 0.01) and `hom-either` (p < 0.03), but not `hom-both` (p = 0.4). However, the AUCs of the two pretrained SSL models are not significantly different for Hausa. Therefore, looking at a more restricted set of consonants which share key phonological features reveals that the multilingual model is significantly less predictive for Shilluk, but not significantly better for Hausa, either.

5 Comparison: Features versus surface representations

In the first two studies, we have seen that acoustic similarity measures and featural similarity measures are better than chance at classifying consonant pairs as participants or non-participants in certain conditions. Study 1 (Section 2) showed that the Count-Sim featural similarity measure predicts participation in consonant harmony only for the `all` consonant pair subgrouping but not the fairest pair comparison. Meanwhile, the second study found that the similarity in surface properties generated by the monolingual SSL model is predictive of participation for all consonant pair subgroupings. Are these differences significant? In other words, is one measure of similarity more predictive of consonant harmony than the other?

Once again, we used DeLong's test to perform this comparison. We focus on evaluating the monolingual model versus the Class-Sim measure as these were the most predictive types of similarity in the previous studies. With Shilluk, the AUCs for monolingual SSL model are not significantly different ($p > 0.05$) than the AUCs for the Class-Sim measure. Interestingly, for Hausa, the AUCs for the monolingual model are significantly above the AUCs for the Class-Sim measure for `hom` ($p < 0.02$), `hom-either` ($p < 0.01$), and `hom-both` ($p < 0.001$), although not for `all` ($p > 0.05$). Taken together, these results show that the Similarity Hypothesis is true for Hausa only with the similarity measure computed with the multilingual SSL model. For Shilluk, the best SSL similarity measure and the best featural similarity measures do not make significantly different predictions for participation in consonant harmony, i.e. the difference in AUCs is not significant. There is thus no measure of similarity which better across the board for the languages examined.

6 Conclusion

Throughout, we have shown that the pairs of consonants in the harmonic classes for both Shilluk and Hausa are statistically more likely to be similar than pairs in the disharmonic classes. In other words, the Similarity Hypothesis is true. That said, there is no single similarity measure which predicts the harmonic classes in both case studies. While the harmonic class in Shilluk is best predicted by the similarity metric over phonological features (Class-Sim), it is cosine similarity in SSL representations which is most predictive of Hausa harmonic classes. Even though Class-Sim has the highest AUC and is thus the best predictor of participation in consonant harmony, it is not significant at the most restrictive subgroupings of consonant pairs, i.e. `hom-either` and `hom-both`.

We have confirmed that the harmonic classes of consonants (i.e. the consonants which are required to agree) in Shilluk and Hausa are well defined as being highly similar on top of homorganic and sharing values for either [SONORANT] or [CONTINUANT] or both. In this sense, the studies above provide a systematic test of the consensus in the literature that harmonizing consonants are similar and exactly what similarity must mean for this to be true (Rose and Walker, 2004; Gallagher, 2010a,b; Hansson, 2010; Mackenzie, 2011, 2016; Arsenault, 2012). What we have also shown is that a universally predictive measure of similarity remains elusive. Assuming that similarity predicts which consonants are required to agree in all languages with consonant harmony, then we still do not have the right definition of similarity.

References

Ahn, Emily, and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5286–5294. URL `https://aclanthology.org/2022.lrec-1.566.pdf`.

Ardila, R., M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4211–4215.

Arsenault, Paul Edmond. 2012. Retroflex Consonant Harmony in South Asia.

Babu, Arun, 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. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. URL `http://arxiv.org/abs/2111.09296`.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. URL `http://arxiv.org/abs/2006.11477`.

Chernyak, Bronya R., Ann R. Bradlow, Joseph Keshet, and Matthew Goldrick. 2024. A perceptual similarity space for speech based on self-supervised speech representations 155:3915–3929. URL `https://pubs.aip.org/jasa/article/155/6/3915/3299166/A-perceptual-similarity-space-for-speech-based-on`.

Cho, Cheol Jun, Peter Wu, Abdelrahman Mohamed, and Gopala K. Anumanchipalli. 2023. Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE. URL `https://ieeexplore.ieee.org/document/10094711/`.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach 44:837. URL `https://www.jstor.org/stable/2531595?origin=crossref`.

Doucette, Amanda, Timothy J O'Donnell, Morgan Sonderegger, and Heather Goad. 2024. Investigating the universality of consonant and vowel co-occurrence restrictions 9. URL `https://www.glossa-journal.org/article/id/9373/`.

Dresher, B Elan. 2003. Contrast and asymmetries in inventories. In *Linguistik Aktuell/Linguistics Today*, ed. Anna Maria Di Sciullo, volume 58, 239–257. John Benjamins Publishing Company. URL `https://benjamins.com/catalog/la.58.10dre`.

Dresher, Bezalel E. 2009. *The contrastive hierarchy in phonology*. Number 121 in Cambridge Studies in Linguistics. Cambridge University Press.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity Avoidance and the OCP 22:179–228. URL `http://link.springer.com/10.1023/B:NALA.0000005557.78535.3c`.

Gafos, Adamantios I. 1996. The Articulatory Basis of Locality in Phonology.

Gallagher, Gillian. 2010a. The perceptual basis of long-distance laryngeal restrictions.

Gallagher, Gillian. 2010b. Perceptual distinctness and long-distance laryngeal restrictions 27:435–480. URL `https://www.cambridge.org/core/product/identifier/S0952675710000217/type/journal_article`.

Gilley, Leoma G. 1992. *An autosegmental approach to Shilluk phonology*. Number publication 103 in Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics. The Summer Inst. of Linguistics [u.a.].

Gwado Ayoker, Otto, and Bert Remijsen. 2017. Collection of Shilluk narratives and songs. URL `https://datashare.is.ed.ac.uk/handle/10283/425`.

Hansson, Gunnar Ólafur. 2010. *Consonant harmony: Long-distance interaction in phonology*. Number v. 145 in University of California Publications in Linguistics. University of California Press.

Hsu, Wei-Ning, 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. URL `http://arxiv.org/abs/2106.07447`.

Mackenzie, Sara. 2011. Contrast and the evaluation of similarity: Evidence from consonant harmony 121:1401–1423. URL `https://linkinghub.elsevier.com/retrieve/pii/S0024384111000477`.

Mackenzie, Sara. 2016. Consonant harmony in Nilotic: Contrastive specifications and Stratal OT 1. URL `https://www.glossa-journal.org/article/id/4823/`.

Martin, Kinan, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration. URL `http://arxiv.org/abs/2306.06232`.

Mortensen, David R, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3475–3484.

Newman, Paul. 2000. *The Hausa Language: An Encyclopedic Reference Grammar*, volume 122.

Pasad, Ankita, Bowen Shi, and Karen Livescu. 2023. Comparative Layer-Wise Analysis of Self-Supervised Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE. URL `https://ieeexplore.ieee.org/document/10096149/`.

Pierrehumbert, Janet B. 1993. Dissimilarity in the Arabic verbal roots 23:367–381.

R Core Team. 2025. *R: A language and environment for statistical computing*. URL `https://www.R-project.org/`.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves 12:77. URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77`.

Rodriguez, Joselyn, Kamala Sreepada, Ruolan Leslie Famularo, Sharon Goldwater, and Naomi Feldman. 2024. Self-supervised speech representations display some human-like cross-linguistic perceptual abilities. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, 458–463. Association for Computational Linguistics. URL `https://aclanthology.org/2024.conll-1.35`.

Rose, Sharon, and Rachel Walker. 2004. A Typology of Consonant Agreement as Correspondence 80:475–531. URL `https://muse.jhu.edu/article/173138`.