What Matters in Tonotactic Learning

Han Li and Jeffrey Heinz
Stony Brook University
Department of Linguistics
Institute for Advanced Computational Science
{han.li.4,jeffrey.heinz}@stonybrook.edu

Abstract

This paper investigates whether tonotactic learning differs across representations and learning models. We conduct an experiment using the same dataset encoded in three representations: segments, features, and autosegmental representations (ARs). To the extent possible, two learning models are evaluated, the Maximum Entropy (MaxEnt) model and the Bottom-Up Factor Inference Algorithm (BUFIA), to examine how learning outcomes interact with both model type and representations. A follow-up experiment further explores the roles of frequency and complexity thresholds. The results show that (1) AR-based learning gives the strongest overall performance; (2) there is no consistent advantage between segmental and feature-based representations across learning models; (3) MaxEnt performance improves substantially when frequency information is introduced and lastly (4) the effects of complexity bounds interact with representation type and frequency information. These findings suggest that tonotactic learning benefits from structurally explicit representations. Overall this work highlights the importance of using linguistically meaningful representations in learning.

1 Introduction

This paper focuses on the learnability of tonal phonotactics, or tonotactics, which addresses the question of how tones and tone-bearing units (TBUs, such as syllable or mora) are organized in tonal languages. More specifically, we examine how the learning outcomes are influenced by representations and learning models.

For a long time, despite the widespread distribution of tonal languages across the world (Yip, 2002), tonotactics has received considerably less attention in computational learning than segmental phonotactics. One possible reason for this gap lies in representational differences. Segmental phonotactics is typically modeled using linear representations such as strings, which makes it readily compatible with a wide range of existing learning models. (Coleman and Pierrehumbert, 1997; Hayes and Wilson, 2008; Goldsmith and Riggle, 2012; Mayer and Nelson, 2020; Schrimpf and Jarosz, 2014). Tonotactics, by contrast, has long been analyzed over non-linear structures in linguistics, most widely autosegmental representations (ARs, Goldsmith, 1976). These more complex representations pose challenges for many existing computational models, which are unable to operate directly over non-linear structures. As a workaround, many approaches usually flatten tonal patterns into strings to fit linear learning frameworks (Shih and Inkelas, 2013). Although a substantial body of work has argued that linear representations are inadequate for capturing tonal generalizations (Clements and Goldsmith, 1984; Odden, 2013) and examined the computational properties of ARs (Bird and Klein, 1990; Coleman and Local, 1991; Kornai, 2007; Jardine and Heinz, 2015; Jardine, 2016, 2017, 2019), we are unaware of any empirical demonstrations of clear advantages for non-linear representations over strings in tonotactic learning.

To fill this gap, we conducted computational experiments and encoded the same dataset in three different representational systems commonly used in tonal phonology, namely strings, sequences of features, and ARs. The string (or segmental) representation uses a sequence of symbols representing low (L), high (H), rising (R), and falling (F) tones. The feature-based representation encodes the data as a sequence of feature bundles with the features [±high] and [±contour]. The third representation provides ARs, in which tones and tone-bearing units are represented on separate tiers linked by association relations (see §2).

These encodings are grounded in theoretical phonology, and differ in the degree to which tone is independent from the TBUs. Since this study fo-

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 180–190

cuses on the effect of representational differences on phonotactic learning, we assume that representations are given to the learner rather than learned from the data. Different representations are evaluated using two learning models: the Bottom-Up Factor Inference Algorithm (BUFIA; Chandlee et al., 2019) and the Maximum Entropy phonotactic learner (MaxEnt; Hayes and Wilson, 2008). BUFIA is a categorical learner that induces a set of inviolable constraints, whereas MaxEnt is a gradient learner that assigns probabilistic weights to constraints. In addition, a set of linguistically motivated grammars, encoded in the corresponding representations, serves as a baseline for comparison with the learned grammars.

One limitation of this study is that the current implementation of MaxEnt does not support ARs. A more comprehensive study would also include a version of MaxEnt which operates over ARs. In addition, neural models have been proposed for learning phonotactic or stress patterns (Goldsmith, 1994; Mayer and Nelson, 2020; Prickett and Pater, 2025; Gupta and Touretzky, 1991), which are not studied here. Furthermore, we do not address the problem of representational parsing or acquisition.[1] We leave these questions for future work. Despite these limitations, the results presented here help assess the extent to which ARs facilitate learning.

The corpus in this study comes from the Hausa database in the *World Loanword Database* (Awagana et al., 2009), which was curated by Li (2025). The corpus contains 621 mono-morphemic, non-borrowed words (64 types) from Hausa's core vocabulary, transcribed in orthography.

This corpus was randomly divided into four partitions (folds). Each fold was converted into the three representational schemes. The tonal representations studied here abstract away from segmental content and retain only tonal and TBU information. For each representation and model, cross-fold validation was conducted and their model performance was examined. In every case, the evaluation set was augmented to include forms that violated the constraints in the baseline grammar, and precision, recall, and F1 scores were reported. These measures help identify whether the models generalize appropriately without overgeneralizing.

The results show three main findings. First, segmental and feature-based representations do not show a consistent difference in learning outcomes: they perform more or less identically under BUFIA and only show marginal and inconsistent differences under MaxEnt. Second, BUFIA over ARs achieves the best performance. Third, learning outcomes are influenced by other factors such as frequency information, complexity bounds, and evaluation criteria. Frequency substantially improves MaxEnt performance and gives comparable results to BUFIA-AR, while categorical learners such as BUFIA remain stable without frequency.

Overall, this work highlights the importance of incorporating linguistically meaningful representations into computational models of phonological learning. It further suggests these approaches should be evaluated on larger-scale and more complex datasets to better identify differences between the approaches.

2 Background

This section provides the theoretical and computational background and methods for the different representations investigated here. We adopt a model-theoretic perspective that treats all representations as different types of mathematical objects (Enderton, 2001), or more specifically, relational structures (Rogers et al., 2013; Chandlee et al., 2019; Rogers and Lambert, 2019).

Model-theoretic representations not only provide a unified framework for encoding different representational systems, but also make explicit how linguistic units are structurally related to one another. One can observe that the more enriched representations encode more complex dependency relations, which in turn lead to different generalizations for phonotactic learners.

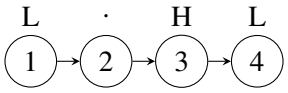
2.1 Representations

We define a phonological representation as relational structures of the form $\langle D, \mathcal{R} \rangle$, where D is a finite domain of positions, and $\mathcal{R}$ is a set of relations (also known as the model *signature*) that connect the composing units of the phonological representation (Strother-Garcia et al., 2017; Jardine, 2016). Relations may be unary, indicating the identity of an element, or binary, specifying the relationship between pairs of elements.

The string representation arranges symbols along a single axis. The signature of a string model includes unary labeling relations indicating

[1] Though see Jardine and Heinz (2015) for one method for deriving ARs from strings.

the symbols, and a single binary successor relation $\rightarrow$ defining linear order. For example, the string representation of the tonal structure of the Hausa word *mù.tûm* ("person") is L.HL. It is visualized in Figure 1. The domain $D = \{1, 2, 3, 4\}$ specifies four positions, which are labeled by their tone and syllable the boundary marker (.). The successor relation encodes the temporal order. Each position occupied by a tone can be thought of representing a single mora. Thus bimoraic syllables will contain a sequence of two tones. For complex tonal forms where a single mora is associated with contour tones, the symbols R (representing rising LH on one mora) and F (representing falling HL on one mora) are used.

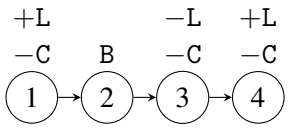


$$
\begin{aligned}
D &= \{1, 2, 3, 4\} \\
L &= \{1, 4\} \\
H &= \{3\} \\
. &= \{2\} \\
R &= \varnothing \\
F &= \varnothing \\
\rightarrow &= \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 3, 4\rangle\}
\end{aligned}
$$

Figure 1: String model when $\Sigma = \{H, L, R, F, .\}$ for the Hausa word *mù.tûm* ("person")

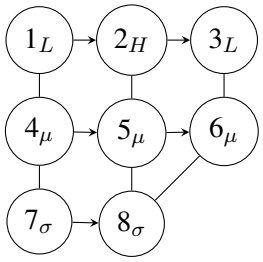
The feature-based representation also arranges symbols along a single axis. Like the string model, the signature of the feature model includes single binary successor relation $\rightarrow$ defining linear order. Unlike the string models, there are five unary relations: `{+Low, -Low, +Contour, -Contour, Boundary}`, where `low` and `contour` are binary features, and `Boundary` is treated as a privative feature. In the string model, each position is true of exactly one unary relation. However, in the feature model, positions may be true of more than one unary relation. Strother-Garcia et al. (2017) and Vu et al. (2018) refer to such models as *unconventional* because they are not the standard way to model strings in theoretical computer science (though it is standard in phonology). Note in this representation `[±Contour]` is only used to indicate monomoraic contour tones. Figure 2 shows the feature representation of same tonal structure of the word in Figure 1.

ARs, in contrast, correspond to multi-linear (tiered) relational structures. The tiers correspond to units such as tones, moras, and syllables. The unary relations are {L, H, μ, σ}, denoting low and high tones, moras and syllables respectively. The signature not only includes the successor relation, which relates positions on the same tier, but it also includes a binary association relation α. This relation captures associations across tiers. In this representation, structural dependencies, such as tone spreading or contour formation, are encoded directly as relations rather than inferred from symbol sequences. For example, Figure 3 shows the AR model for the same word *mù.tûm*. The model consists of eight units, defined by $D = \{1, \ldots, 8\}$, arranged across three tiers, each following a successor relation $\rightarrow = \{(1, 2), (2, 3), \ldots\}$. The association relations, on the other hand, are expressed as a set of unordered pairs $\alpha = \{(1, 4), (2, 5), \ldots, (6, 8)\}$, indicating which two elements are connected by an association line.



$$
\begin{aligned}
D &= \{1, 2, 3, 4\} \\
\texttt{+Low} &= \{1, 4\} \\
\texttt{−Low} &= \{3\} \\
\texttt{+Contour} &= \varnothing \\
\texttt{−Contour} &= \{1, 3, 4\} \\
\texttt{Boundary} &= \{2\} \\
\rightarrow &= \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 3, 4\rangle\}
\end{aligned}
$$

Figure 2: Feature model for *mù.tûm* ("person")



$$
\begin{aligned}
D &= \{1, 2, 3, 4, 5, 6, 7, 8\} \\
L &= \{1, 3\} \quad H = \{2\} \quad \mu = \{4, 5, 6\} \quad \sigma = \{7, 8\} \\
\rightarrow &= \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 4, 5\rangle, \langle 5, 6\rangle, \langle 7, 8\rangle\} \\
\alpha &= \{(1, 4), (2, 5), (3, 6), (4, 7), (5, 8), (6, 8)\}
\end{aligned}
$$

Figure 3: Autosegmental model for *mù.tûm* ("person")

This formalization explicitly captures how within-tier and across-tier relations are represented in autosegmental phonology (Goldsmith,

1976). In particular, tonal generalizations have been argued to be computationally simpler over ARs than over strings, because cross-tier dependencies are represented directly as relations. This allows processes to be modeled by modifying associations rather than altering feature values or symbols, and can describe long-distance processes locally (Goldsmith, 1976; Jardine, 2017).

2.2 Phonotactic Learning Models

One learning system used in this paper is the BUFIA (Chandlee et al., 2019) which learns constraints expressed as relational structures under an arbitrary model-theoretic signature. This means the general form of the algorithm can be flexibly applied to very many specific linguistic representations. In this work, we used three versions of BUFIA: BUFIA-Seg, BUFIA-Ftr and BUFIA-AR.

In a nutshell, BUFIA is a deterministic phonotactic learner that traverses a hypothesis space of all logically possible structures, finding those relational structures which are absent in the data. BUFIA constructs the hypothesis space on the fly, beginning from the simplest (smallest) relational structures, and considering increasingly complex ones as needed. The notions 'simpler' and 'more complex' are made concrete by the *containment relation*: some pairs of structures stand in a containment relation (one containing the other, where the more complex structure is referred to as the *superfactor* and the contained structure is the *subfactor*), while others are not comparable. All logically possible structures, which serve as candidate constraints, are partially ordered by the containment relation.

Two core operations of BUFIA determine its search: containment and the sub/superfactor relation. The learner is provided with a set of positive data and a complexity bound. The search begins from an initial structure at the bottom of hypothesis space (typically the empty structure). For each candidate, BUFIA checks whether it is contained in the positive dataset. If it is attested, the learner generates its *next superfactors* (i.e., minimally more complex superfactors) and continues searching upward. Otherwise, the structure is identified as a forbidden factor (i.e., a *constraint*), and its superfactors are pruned from the rest of the space. Chandlee et al. (2019) prove that BUFIA identifies the most general constraints within the complexity bounds that are consistent with the data. The resulting grammar guarantees full coverage of the observed data while remaining maximally general.

As said before, a key merit of BUFIA is its representational flexibility. The algorithm can operate over any representational choices, provided they are defined model-theoretically. Recent work has demonstrated the success of BUFIA in phonotactic learning. Swanson et al. (2026) implemented a string-based version of BUFIA for learning Quechua phonotactics and showed that it performed as well as the (Hayes and Wilson, 2008; Wilson and Gallagher, 2018). Li (2025) introduced a version of BUFIA that operates directly over ARs for tonotactic learning and showed that it could learn tonotactic constraints. This representational flexibility allows controlled comparison of different representations of the same data under a shared learning framework.

BUFIA is a categorical learner: a candidate structure is either well-formed or ill-formed, depending on whether it contains any forbidden subfactor. In contrast, probabilistic models, such as MaxEnt (Hayes and Wilson, 2008), assign gradient well-formedness values to candidate forms. Well-formedness is interpreted as a probability rather than a binary value; consequently, constraints are treated as violable, each associated with a penalty weight. This evaluation is cumulative: the more constraints a candidate violates, and the higher the weights of those constraints, the lower its predicted probability of well-formedness.

MaxEnt has been widely used in string-based phonotactic learning, and more recent implementations induce phonological tiers to capture long-distance patterns (Gouskova and Gallagher, 2020) and metrical structure (Lee et al., 2025). These studies suggest that probabilistic string-based approaches can learn structural information that is not directly visible in surface sequences. However, these representations differ from ARs in that they typically do not encode association relations between individual elements on separate tiers.

The central question of this paper is the extent to which representations facilitate tonotactic learning. Does adopting ARs improve learning outcomes over adopting string and featural representations? A related question is whether learning outcomes are influenced by the type of learner (here BUFIA or MaxEnt). In the following section, we present an experiment designed to test these questions by applying the same dataset across different representations and the two learning schemes.

3 Experiment: Hausa Tonotactics

To test tonotactic learnability across representations and models, we conducted a series of controlled computational experiments on the same dataset. This section details the experimental setup, including construction of the training and test data, and the evaluation metrics used to compare model performance.

3.1 Hausa Data

The language data used in this study come from Hausa, a major tonal language spoken in West Africa. Hausa has three surface tones in orthography: H, L, and HL (a falling tone, which occurs only in heavy syllables). Previously proposed generalizations about Hausa tonotactics are listed in Table 1 (Newman, 2002; Leben, 1996; Zoll, 2003).

Constraints	Interpretation
(1) ***Rise**$_\sigma$	No monosyllabic LH
(2) ***Contour**$_\mu$	No contour on single mora
(3) ***3T**	No trimoraic contour
(4) ***NonFinHL**	No non-final HL
(5) ***Lapse**	No adjacent L-toned syllables

Table 1: Phonological Generalizations in Hausa

We use a curated corpus from Li (2025), constructed from the Hausa database (Awagana et al., 2009) in the *World Loanword Database*. The corpus contains 621 mono-morphemic, non-borrowed words from the core vocabulary, transcribed in orthography. These 621 words yield a total of 64 tonal structures.

Among the constraints in Table 1, the first three constraints, *Rise$_\sigma$,*Contour$_\mu$,*3T, are hard constraints that define the basic phonology of tones and syllable structures of Hausa. They are never violated in the corpus. Constraint *NonFinHL applies to the non-derived Hausa words where the falling contour (HL) can only occur word finally. The constraint *Lapse is proposed by Zoll (2003) to explain the rarity of adjacent L-toned syllables. Newman (2002) points out that there are exceptional words that violate *NonFinHL and *Lapse. Li (2025) also reports that the corpus itself includes forms violating these two constraints.

3.2 Procedure

The experiment compares grammars learned under multiple conditions, varying both learning model (BUFIA vs. MaxEnt) and representation (Segment, Feature, AR). These learned grammars are evaluated against baseline grammars defined over the corresponding representations (Baseline-Seg/Ftr/AR, see Appendix A1). Table 2 summarizes all model conditions. As mentioned, MaxEnt-AR is one model we are unable to test.

Model	length
Baseline-Seg	$k = 3$
Baseline-Ftr	$k = 3$
Baseline-AR	$t = 2, s = 2, m = 3$
BUFIA-Seg	$k = 3$
BUFIA-Ftr	$k = 3$
BUFIA-AR	$t = 2, s = 2, m = 3$
MaxEnt-Seg	$k = 3$
MaxEnt-Ftr	$k = 3$

Table 2: Learning models in the experiment

Both BUFIA and MaxEnt specify thresholds that limit the maximum length or structural complexity of learned constraints. These can be set manually, or "tuned" according to data. However, the data used for tuning cannot be used for training or testing. Since our corpus is small (64 forms), we elected to manually set these thresholds. We chose to set values according to the constraints in the baseline grammar, which helps provide a fair comparison. Thus, we set the thresholds to the most complex constraint in the baseline grammar. For string and featural representations, the most complex baseline constraint has length 3 (e.g., *HL.). For ARs, the threshold consists of three parameters: t (maximum number of tonal units), s (maximum number of syllables), and m (maximum number of moras). Since the most complex constraint in the baseline grammar has two tones, two syllables and three moras, we set $t = 2$, $s = 2$, and $m = 3$.

We evaluated each model using 4-fold cross-validation. The dataset was partitioned into four equally sized folds (16 forms per fold). In each iteration, three folds were used for training and the remaining fold provided the positive items in the test set. This process was repeated four times, with each fold serving as the test set once.

Arguably, if the test set only contains positive items, then it can mislead. For example, a grammar with no constraints will accept the entire test set because no test form violates any constraint. Such a grammar therefore over-generates. The in-

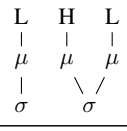
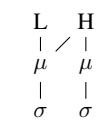
Word	String	Feature	AR
mù.tûm ‘person’	L.HL	L: +0−+ C: −0−− B: 0+0 0	
(nonce)	R.H	L: +0− C: +0− B: 0+0	

Table 3: Positive and negative forms in three representations

clusion of negative forms in the test set helps detect over-generation. While the grammar with no constraints would correctly classify all positive forms, it would also incorrectly classify all negative forms. Similarly, a grammar with too many constraints may correctly classify all negative forms, and incorrectly reject all positive forms. Grammars that more correctly classify both positive and negative forms are desirable.

Therefore, we also constructed negative items to include in the test set. Since there were 16 positive items in each test set, we constructed 16 negative items, and added them to each test set. Thus, each test set contained 32 forms.

In the absence of native speaker judgments, there is less certainty regarding whether an unattested form is negative or positive. For example, the baseline grammar includes two constraints that are known to have exceptions.

We therefore constructed negative forms that met two conditions. First, each negative form was unattested in the corpus. Second, each negative form violated at least one of the hard constraints in Table 1 (constraints 1-3). Since there were 16 negative forms and three constraints, at least five forms violated each constraint. One positive and one negative form are shown in Table 3.

For training and testing each model, all data was converted into each of the three representational systems: string, feature, and ARs. To evaluate model performance, we measured learning outcomes as a binary classification task.

3.3 Evaluation

For grammars returned by BUFIA and the baseline grammars, if a test form violated any of the constraints in the grammar, then the grammar marked it as rejected; otherwise, it was marked as accepted. A prediction was counted as correct if positive forms were accepted and negative forms were rejected. Performance was measured using averaged recall, precision, and F1 score across the folds.

The MaxEnt model produces a gradient grammar with weighted constraints. This grammar assigns a score to each test form based on a weighted sum of constraint violations. A zero score means the form violates no constraints, and higher scores indicate the form violates multiple, or higher-weighted, constraints.

There are many ways this score can be interpreted to produce a binary classification. We considered a simple threshold. If the score the grammar assigns to a form is greater than the threshold then it is rejected; otherwise it is accepted. We examined a few possible thresholds and selected the one that gave the best results. Again, normally the threshold could be set as part of a tuning phase, but we did not want to spare any data for this.

The three thresholds we considered are the following. First, over-zero: a form is rejected if its total weighted violation score is greater than zero. Second, over-average: a form is rejected if its total weighted violation score exceeds the average violation score across all candidate forms. Third, over-median: a form is rejected if its total weighted violation score exceeds the median score across all candidate forms. These criteria are different ways to categorically interpret the gradient grammar.

In all of our experiments, the over-average interpretation performed the best. The over-zero interpretation was the worst, and tended to reject more than half of the positive forms. This is likely because constraints with small non-zero weights were learned. Generally, the over-median approach performed similarly to over-average, but had lower scores. Consequently, in what follows, the MaxEnt results are reported with the over-average threshold.

4 Results

Results of the experiments are shown in Table 4. Several aspects are worth noting. First, the baseline grammars have identical results. A closer inspection of the errors confirms that they also coincide. This indicates a one-to-one correspondence between representations, and the forms that violate these constraints. It follows that representational differences do not affect how this fixed grammar classifies forms.

On the other hand, for the BUFIA and Max-

Model	Recall	Precision	F1
Base-Seg	0.812	1.000	0.894
Base-Ftr	0.812	1.000	0.894
Base-AR	0.812	1.000	0.894
BUFIA-Seg	0.984	1.000	0.992
BUFIA-Ftr	0.984	1.000	0.992
BUFIA-AR	1.000	1.000	1.000
MaxEnt-Seg	0.922	0.827	0.871
MaxEnt-Ftr	0.906	0.773	0.832

Table 4: Average model performance

Ent models, which learn constraints, performance does vary depending on both the representation and the learning model. Among the BUFIA models, while there is no difference between BUFIA-Seg and BUFIA-Ftr, both achieve an F1 score of 0.992. BUFIA-AR, however, performs perfectly (F1 = 1.000). Although it is true that BUFIA-AR outperforms BUFIA-Seg and BUFIA-Ftr, all the BUFIA models are exhibiting ceiling levels of performance, or are extremely close to it.

Choice of representation also impacts the MaxEnt models. MaxEnt-Ftr achieves an F1 score of 0.832, while MaxEnt-Seg reaches 0.871. The better performance on segmental representations over featural ones contrasts with earlier studies with MaxEnt on segmental phonotactics, which found that learned constraints over features were more effective than learned constraints over segments (Hayes and Wilson, 2008; Wilson and Gallagher, 2018).

Finally, the results show that BUFIA-Seg and BUFIA-Ftr outperform MaxEnt-Seg and MaxEnt-Ftr. We were surprised by this result and conducted follow-up experiments to see if we could improve the performance of MaxEnt.

One finding is that ARs provide a learning advantage over segmental or feature-based representations. However, the improvement is marginal as all the BUFIA models performed at, or close to, ceiling. We therefore conclude that while there is some evidence that ARs improve tonotactic learnability, we believe it is premature to draw firm conclusions. The tonotactics of Hausa may simply be too simple to see any significant difference in BUFIA's learning capacities with different representations. Therefore, we think that similar experiments ought to be conducted on languages with more complex tonotactics as that may allow for clearer distinctions to emerge when learning over segmental, featural, or autosegmental representations.

5 Follow-Up Experiments

We were curious what hindered the performance of the MaxEnt model. In the previous experiments, the model was trained on the types rather than tokens; therefore, frequency information was not available during learning. Since MaxEnt determines weights for the constraints it learns partially based on frequencies, we hypothesized their absence may have been a hindrance.

Therefore, in a follow-up experiment, we introduced frequency to MaxEnt by training the model on token data rather than type data (621 tokens in total). As in the previous setup, we conducted cross-fold validation over four folds (each fold contained 156 or 155 tokens). As before, an equal number of negative forms were included in the test set. Since the over-mean criterion gave the best results in the previous experiment, we use it again here to calculate recall, precision, and F1 scores.

Another question is how to determine an appropriate complexity bound that accounts for the data without leading to over-generalization. To examine this, we also ran experiments with MaxEnt models where the maximum constraint length was set to 4. This is the highest value permitted by the implementation we used. We conducted experiments where these MaxEnt models had access to frequency information, and where they did not.

As shown in Table 5, the inclusion of frequency information makes a substantial difference. When frequency information was included, all MaxEnt models performed close to ceiling with F1 scores greater than or equal to 0.990 regardless of the complexity bound. Without frequency information, the highest F1 score, achieved by MaxEnt-Seg, is 0.871 when $k = 3$, and the other MaxEnt models get even lower when k increases to 4.

In terms of the feature vs. segment comparison, the results are inconsistent. When frequency information is not included, the advantage between MaxEnt-Seg and MaxEnt-Ftr reverses between complexity bounds: MaxEnt-Seg performs better at $k = 3$, whereas MaxEnt-Ftr performs better at $k = 4$. Once the frequency information is introduced, there appears to be no appreciable difference among the F1 scores on the grammars obtained by the MaxEnt models between constraint

Model	k	Recall	Precision	F1
MaxEnt-Seg	3	0.922	0.827	0.871
MaxEnt-Ftr	3	0.906	0.773	0.832
MaxEnt-Seg-freq	3	1.000	0.981	0.990
MaxEnt-Ftr-freq	3	1.000	0.981	0.990
MaxEnt-Seg	4	0.644	0.866	0.651
MaxEnt-Ftr	4	0.816	0.844	0.829
MaxEnt-Seg-freq	4	1.000	0.994	0.997
MaxEnt-Ftr-freq	4	1.000	0.980	0.990
BUFIA-Seg/Ftr	3	0.985	1.000	0.992
BUFIA-Seg/Ftr	4	0.969	1.000	0.984
BUFIA-Seg/Ftr	5	0.907	1.000	0.950
BUFIA-Seg/Ftr	6	0.782	1.000	0.873
BUFIA-Seg/Ftr	7	0.656	1.000	0.782

Table 5: Performance of MaxEnt and BUFIA models across complexity bounds and frequency settings. The MaxEnt-Seg and MaxEnt-Ftr results for $k = 3$ are repeated from Table 2 for ease of comparison.

lengths 3 and 4, as these results are effectively at ceiling.

For the sake of comparison, we also examined the grammars returned by the BUFIA models when we varied the maximum sizes of the constraints. For BUFIA-Seg and BUFIA-Ftr, we examined constraint lengths from 3 to 7, as shown in Table 5; For BUFIA-AR, we explored parameter settings for t, m, s up to 3. The results for $m = 3$ are presented in Table 6, while the complete set of results is provided in the Appendix A2.

For BUFIA-Seg and BUFIA-Ftr, the highest F1 score is 0.992 when $k = 3$. For $k = 4$, the F1 is still close to ceiling at 0.984. As k increases, the F1 continues to drop for these models, indicating over-fitting.

For BUFIA-AR, when $m < 3$, the highest F1 score was 0.842. When the search for constraints cannot consider structures with three moras, BUFIA-AR will not find the constraint against trimoraic contours. Similarly, when $t = 1$, the hard constraints against monosyllabic rises and monomoraic contours cannot be learned, which significantly brought down performance. On the other hand, F1 scores were at 1.000 in the following three settings: $[m = 3, s = 1, t = 2], [m = 3, s = 1, t = 3], [m = 3, s = 2, t = 2]$. In the maximum setting when $[m = 3, s = 3, t = 3]$, the F1 score dropped to 0.967, indicating the beginning of the cline towards over-fitting.

m	s	t	Recall	Precision	F1
3	1	1	1.000	0.615	0.762
3	1	2	1.000	1.000	1.000
3	1	3	1.000	1.000	1.000
3	2	1	1.000	0.615	0.762
3	2	2	1.000	1.000	1.000
3	2	3	0.953	1.000	0.975
3	3	1	0.984	0.612	0.754
3	3	2	0.984	1.000	0.992
3	3	3	0.938	1.000	0.967

Table 6: Performance of BUFIA-AR across complexity thresholds when $m = 3$.

6 Discussion and Conclusion

We observe from the results that several factors influence learning outcomes in surface-based tonotactics.

First, Wilson and Gallagher (2018) show that for segmental phonotactics, learning tends to be more successful over feature-based representations than over segments. Our results in tonal phonology, however, do not show a consistent difference between featural and segmental representations. Under the BUFIA model, the two representations yield identical performance, while under MaxEnt the difference is marginal and inconsistent. One possible explanation is that string-based tonal representations are relatively simple, involving a much smaller inventory of tonal symbols (typically two to four tonemes and three to six features) compared with the richer feature systems used in segmental phonology.

In contrast, ARs can be used effectively for learning and consistently outperform string-based representations. This advantage is particularly evident when compared with MaxEnt models without frequency, and the improved performance is marginal when compared with MaxEnt models trained with frequency.

In addition, for gradient learners, frequency information in the dataset is crucial for achieving strong performance; however, this is not required for BUFIA, which achieves competitive results even without frequency information. Furthermore, for a dataset such as Hausa, which exhibits a relatively small set of surface tonal constraints, the optimal complexity threshold appears to lie between $k = 3$ and $k = 5$ for string representations, or under AR settings where $\mu \geq 3$ and $t > 1$. We be-

lieve for more complex tonal systems, learners like BUFIA-AR are advantageous as they allow complexity to be controlled more finely across multiple tiers while searching the hypothesis space.

We aim to draw attention to the use of computational methods for evaluating competing representational choices and theories (Oakden, 2020; Danis, 2025). Models such as BUFIA, which allow flexible representations, can serve as practical tools for facilitating such comparisons. More broadly, computational approaches also allow us to explore how grammar inference can be improved through interactions among representations, token or type data, complexity bounds, and evaluation criteria.

In conclusion, we show that tonotactics can be learned effectively and efficiently over autosegmental representations using BUFIA-AR with sparse and limited data. The model achieves a comparable learning outcome to gradient models that rely heavily on distributional information. The results suggest that AR-based learning is a promising direction for modeling tonal grammars and tonal processes, where phonological changes often involve modifications of associations between tiers rather than individual symbols.

This work remains limited in some respects that motivate future research. First, we conducted the experiments on a small dataset with comparatively simple tonotactic patterns, and therefore the performance across learners does not differ greatly. Future work should evaluate these approaches on larger-scale and more complex datasets. This would enable more statistically meaningful comparisons and experiments on probabilistic or neural phonotactic learners. Nonetheless, we hope that by making non-linear representations available for learning, it will become easier to investigate representational learning itself or phonological transformations in the future.

Acknowledgments

HL would like to express her gratitude to John Goldsmith, Ellen Broselow, Thomas Graf, as well as the audience at the Workshop on The Role of Representation in Computational Phonology, for their valuable feedback and support. HL also gratefully acknowledges the support of the Institute for Advanced Computational Science through the Junior Researcher Award. This research was also supported by US NSF grant #2416183 to JH. The authors would also like to thank the four anonymous reviewers for their thoughtful feedback.

References

Ari Awagana, H. Ekkehard Wolff, and Doris Löhr. 2009. Hausa. In Martin Haspelmath and Uri Tadmor, editors, *World Loanword Database*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Steven Bird and Ewan Klein. 1990. Phonological events. *Journal of Linguistics*, 26(1):33–56.

Jane Chandlee, Rémi Eyraud, Jeffrey Heinz, Adam Jardine, and Jonathan Rawski. 2019. Learning with partially ordered representations. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 91–101.

George N Clements and John Goldsmith. 1984. *Autosegmental studies in Bantu tone*, volume 3. Walter de Gruyter.

J. S. Coleman and J. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pages 49–56. Somerset, NJ: Association for Computational Linguistics.

John Coleman and John Local. 1991. The "No Crossing Constraint" in Autosegmental phonology. *Linguistics and Philosophy*, pages 295–338.

Nick Danis. 2025. Logical transductions are not sufficient for notational equivalence. In *Proceedings of the Annual Meetings on Phonology*, volume 1. University of Massachusetts Amherst Libraries.

Herbert B Enderton. 2001. *A Mathematical Introduction to Logic*. Elsevier.

John Goldsmith. 1994. A dynamic computational theory of accent systems. In Jennifer Cole and Charles Kisseberth, editors, *Perspectives in Phonology*, pages 1–28. Stanford: Center for the Study of Language and Information.

John Goldsmith and Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

John Anton Goldsmith. 1976. *Autosegmental Phonology*. Ph.D. thesis, MIT.

Maria Gouskova and Gillian Gallagher. 2020. Inducing nonlocal constraints from baseline phonotactics. *Natural Language and Linguistic Theory*, 38(1):77–116.

Prahlad Gupta and David Touretzky. 1991. What a perceptron reveals about metrical phonology. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pages 334–339.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Adam Jardine. 2016. *Locality and non-linear representations in tonal phonology*. Ph.D. thesis.

Adam Jardine. 2017. The local nature of tone-association patterns. *Phonology*, 34(2):363–384.

Adam Jardine. 2019. The expressivity of autosegmental grammars. *Journal of Logic, Language and Information*, 28:9–54.

Adam Jardine and Jeffrey Heinz. 2015. A concatenation operation to derive autosegmental graphs. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 139–151.

András Kornai. 2007. *Mathematical linguistics*. Springer Science & Business Media.

William R Leben. 1996. Tonal feet and the adaptation of english borrowings into hausa. *Studies in African Linguistics*, 25(2):129–154.

Seung Suk Lee, Joe Pater, and Brandon Prickett. 2025. Representing and learning stress in a maxent framework. In *Proceedings of the Annual Meetings on Phonology*, volume 1. University of Massachusetts Amherst Libraries.

Han Li. 2025. Learning tonotactic patterns over autosegmental representations. In *Proceedings of the Annual Meetings on Phonology*, volume 1. University of Massachusetts Amherst Libraries.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 291–301.

Paul Newman. 2002. *The Hausa Language: An Encyclopedic Reference Grammar*, volume 122. Yale University Press, New Haven, US.

Chris Oakden. 2020. Notational equivalence in tonal geometry. *Phonology*, 37(2):257–296.

David Odden. 2013. *Introducing phonology*. Cambridge University Press.

Brandon Prickett and Joe Pater. 2025. Learning and generalizing stress patterns with a sequence-to-sequence neural network. *Linguistics Vanguard*, (0).

James Rogers, Jeffrey Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert, and Sean Wibel. 2013. Cognitive and sub-regular complexity. *Lecture Notes in Computer Science*.

James Rogers and Dakotah Lambert. 2019. Some classes of sets of structures definable without quantifiers. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 63–77, Toronto, Canada. Association for Computational Linguistics.

Natalie M Schrimpf and Gaja Jarosz. 2014. Comparing models of phonotactics for word segmentation. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 19–28.

Stephanie Shih and Sharon Inkelas. 2013. A subsegmental correspondence approach to contour tone (dis) harmony patterns. In *Proceedings of the annual meetings on phonology*.

Kristina Strother-Garcia, Jerey Heinz, and Hyun Jin Hwangbo. 2017. Using model theory for grammatical inference: a case study from phonology. In *International Conference on Grammatical Inference*, pages 66–78. PMLR.

Logan Swanson, Jeffrey Heinz, and Jon Rawski. 2026. Phonotactic learning with structure, not statistics. *Linguistic Inquiry*. In press.

Mai H Vu, Ashkan Zehfroosh, Kristina Strother-Garcia, Michael Sebok, Jeffrey Heinz, and Herbert G Tanner. 2018. Statistical relational learning with unconventional string models. *Frontiers in Robotics and AI*, 5:76.

Colin Wilson and Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: A case study of south bolivian quechua. *Linguistic Inquiry*, 49(3):610–623.

Moira Yip. 2002. *Tone*. Cambridge University Press.

Cheryl Zoll. 2003. Optimal tone mapping. *Linguistic Inquiry*, 34(2):225–268.

A Appendix

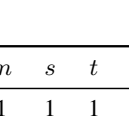
Constraint	String	Feature			AR
***Rise**$_\sigma$	LH	L:	+	−	
		C:	−	−	
		B:	0	0	
***Contour**$_\mu$	R	L:	+		
		C:	+		
		B:	0		
	F	L:	−		
		C:	+		
		B:	0		
***3T**	HHH	L:	−	−	−
		C:	−	−	−
		B:	0	0	0
	LLL	L:	+	+	+
		C:	−	−	−
		B:	0	0	0
	HHL	L:	−	−	+
		C:	−	−	−
		B:	0	0	0
	HLL	L:	−	+	+
		C:	−	−	−
		B:	0	0	0
	HLH	L:	−	+	−
		C:	−	−	−
		B:	0	0	0
***NoFinHL**	HL.	L:	−	+	0
		C:	−	−	0
		B:	0	0	+
***Lapse**	L.L	L:	+	0	+
		C:	−	0	−
		B:	0	+	0

Table A1: Baseline constraints represented in string, feature, and autosegmental representations. Note the string and featural representations for *3T would also include LHH, LHL, and LLH, but these patterns are ruled out by *Rise$_\mu$.

m	s	t	Recall	Precision	F1
1	1	1	1.000	0.500	0.667
1	1	2	1.000	0.615	0.762
1	1	3	1.000	0.615	0.762
1	2	1	1.000	0.500	0.667
1	2	2	1.000	0.615	0.762
1	2	3	1.000	0.615	0.762
1	3	1	1.000	0.500	0.667
1	3	2	1.000	0.615	0.762
1	3	3	1.000	0.615	0.762
2	1	1	1.000	0.500	0.667
2	1	2	1.000	0.727	0.842
2	1	3	1.000	0.727	0.842
2	2	1	1.000	0.500	0.667
2	2	2	1.000	0.727	0.842
2	2	3	0.969	0.721	0.826
2	3	1	1.000	0.500	0.667
2	3	2	1.000	0.727	0.842
2	3	3	0.969	0.721	0.826
3	1	1	1.000	0.615	0.762
3	1	2	1.000	1.000	1.000
3	1	3	1.000	1.000	1.000
3	2	1	1.000	0.615	0.762
3	2	2	1.000	1.000	1.000
3	2	3	0.953	1.000	0.975
3	3	1	0.984	0.612	0.754
3	3	2	0.984	1.000	0.992
3	3	3	0.938	1.000	0.967

Table A2: Performance of BUFIA-AR across complexity bounds (t, s, m).