Using LSTMs to Assess the Obligatoriness of Phonological Distinctive Features for Phonotactic Learning

Nicole Mirea[*] and **Klinton Bicknell**[‡,*]

[*]Northwestern University [‡]Duolingo
nimirea@u.northwestern.edu
klinton@duolingo.com

Abstract

To ascertain the importance of phonetic information in the form of phonological distinctive features for the purpose of segment-level phonotactic acquisition, we compare the performance of two recurrent neural network models of phonotactic learning: one that has access to distinctive features at the start of the learning process, and one that does not. Though the predictions of both models are significantly correlated with human judgments of non-words, the feature-naive model significantly outperforms the feature-aware one in terms of probability assigned to a held-out test set of English words, suggesting that distinctive features are not obligatory for learning phonotactic patterns at the segment level.

1 Introduction

Knowing a language involves having systematic expectations about the sequential sound patterns within syllables and words in the language—a sensitivity to the *phonotactic generalizations* that exist in the language. This sensitivity helps language users segment a continuous stream of speech (Vitevitch et al., 1997), incorporate new words into the lexicon (Storkel et al., 2006), and reconstruct parts of an utterance that may have been obscured by noise. However, the details of how language learners infer these phonotactic generalizations from incoming acoustic data are still unclear. The current project seeks to clarify the extent to which phonetic information (at the level of phonological distinctive features) is useful for predicting upcoming phones within a word, by building computational models of phonotactic acquisition.

Phonotactic patterns are typically stated in terms of generalizations over natural classes; for example, voiced stops cannot follow voiceless stops word-finally in English. These natural classes are defined by a hierarchy or set of distinctive features that is either taken to be universal across languages (Chomsky and Halle, 1965; Clements, 2009) or emergent from the process of phonological acquisition—including phonotactic acquisition—in a particular language (Mielke, 2008; Dresher, 2015). Nevertheless, most models of phonotactic acquisition require that phonological distinctive features be specified in advance of learning. Our work interrogates this assumption through the following questions:

1. Is external information regarding phonological distinctive features a necessary prerequisite for learning word-level phonotactic generalizations?

2. Must models become sensitive to phonological properties of incoming segments in order to represent phonotactic generalizations?

To answer them, we use recurrent neural networks with long short-term memory (LSTM) nodes, which have shown considerable success in learning patterns at the word (Sundermeyer et al., 2015) and character (Kim et al., 2016) levels. These models encode each phonetic segment in the inventory as a vector of numbers. With exposure to more training data, these representations adapt to the task at hand: incrementally predicting each segment in a word, given all previous segments in the word.

If phonetic segments must be specified in terms of distinctive features in advance of phonotactic learning, we would expect a model that encodes phonetic segments in this manner the outset of training to ultimately represent phonotactic generalizations more accurately than one that initially encodes each phonetic segment as a random vector containing no phonetic information whatsoever. If, on the other hand, all information required to

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1595–1605
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

learn phonotactic generalizations is already latent in the sequence of segments, then the featurally-informed model should have no advantage. Alternatively, initializing the model with distinctive features might constrain it to explore a suboptimal area of the solution space, ultimately leading to a less accurate representation of phonotactics.

To investigate our second question, we determine whether the resultant learned encodings of each phonetic segment reflect phonetic information by examining the state of the models after training. If the post-training encodings do encode phonetic information, it would support the centrality of a phonetic representation of incoming acoustic data for phonotactic learning.

Following previous work by Futrell et al. (2017), in Experiment 1 we evaluate how well our models capture phonotactic generalizations by measuring the probability they assign to unseen words from a test corpus. In accordance with rational analysis (Anderson and Milson, 1989), we make claims about the mind by studying the environment in which it operates, under the assumption that the mind adapts to the environment in order to achieve its goals—here, the goal of learning what constitutes a "likely" word-form in a language.[1] If the optimal way of achieving this relies on phonological distinctive features, then we should expect that language users do draw upon this resource in order to infer phonotactic regularities.

To verify this expectation, in Experiment 2 we evaluate our models using the more traditional means of assessing phonotactic learners: comparison with human wordlikeness ratings of non-words. If our models have indeed learned an ecologically valid representation of English phonotactics, we expect the probabilities that they assign to non-words to correlate with wordlikeness ratings assigned by English speakers.

2 Related Work

To ask whether distinctive features are helpful for phonotactic generalization, it is first essential to establish what form these phonotactic generalizations should take. Experimental work supports a characterization of phonotactics as gradient expectations over sequences of sounds, instead of categorical restrictions designating certain sound sequences as marked. In the phonotactic learning experiment conducted by Goldrick (2004), participants were able to acquire feature-based phonotactic constraints of both gradient and categorical forms. Gradient phonotactic sensitivity has also been found in children's productions (Coady and Aslin, 2004) as well as adults' wordlikeness judgments (Frisch et al., 2000). Following this, our model will represent gradient constraints, and its task will be to assign gradient acceptability ratings to sequences of phonetic segments.

Bernard (2017) demonstrated that humans are capable of simultaneously tracking and learning phonotactic generalizations defined at the level of word boundaries, syllable positions, and co-occurrences between adjacent phonetic segments. Our LSTM networks are capable of capturing all three types of constraints. Crucially, they are capable of representing dependencies between non-adjacent units in a sequence (Sundermeyer et al., 2015), which means that they can learn gradient phonotactic constraints at both the word, syllable, and segment level, without the need for explicit syllable coding in the training data.

Many models have addressed the question of how phonotactic generalizations are induced from incoming data (Hayes and Wilson, 2008; Albright, 2009; Futrell et al., 2017)[2]. These vary in terms of the algorithm that the learner uses to learn correspondences between segments. Nevertheless, most of these models of phonotactic acquisition presuppose that incoming data is encoded in terms of a set or hierarchy of distinctive features that are predetermined by the researcher. Our research questions this fundamental assumption, with potential implications for these phonotactic learning models if the assumption is unsubstantiated.

This assumption has already been challenged by a baseline from Albright (2009), which compared bigram models over distinctive features and segments. The segmental bigram model yielded slightly higher agreement with human wordlikeness judgments than the featural bigram model, although the featural bigram model was closer to human judgments for words containing unattested sequences. However, these results may change for models capable of learning generalizations across longer units of structure; this possibility warrants another test.

[1] This goal is subordinate to other goals: speech segmentation, word learning, perception of speech in noise, communication, survival, etc. We have chosen this as a tractable level of analysis.

[2] See Daland et al. (2011) for a comprehensive review.

Previous attempts to explicitly quantify the relevance and "psychological accuracy" of a universal, innate set of distinctive features for phonotactic learning have also produced mixed results. Mielke (2008) used a typological analysis to argue for language-specific, learned distinctive features; in 2012, Mielke devised another phonetic similarity metric that corresponds to surface phonological patterns roughly as well as distinctive features do. Drawing upon this work, Dunbar et al. (2015) compared how well featural representations derived from acoustic, articulatory, and phonotactic models capture phonemic distinctions in English. The phonotactic-derived feature representations performed markedly worse than the acoustic or articulatory representations at separating this phonemic space, suggesting a weaker-than-expected link between phonotactics and acoustic/articulatory phonetics—and indeed, between phonotactics and the features required to distinguish phonemic space. Our work probes this link in the opposite direction, questioning the extent to which distinctive features are necessary to learn phonotactic generalizations.

3 Model[3]

Our models are recurrent neural networks with LSTM nodes. Each network's task is to incrementally predict the next phonetic segment in a sequence, given the beginning of the sequence as input. Models were constructed using PyTorch 0.3.1 (Paszke et al., 2017).

The function and description of each layer in the model is as follows:

3.1 Input Layer

The input layer reads in each phonetic segment is a one-hot vector. The number of nodes in this layer is equal to the size of the phonetic inventory—i.e., the number of unique phones in the corpus (with vowels of different stress levels counted as separate phones). For the present data, this number is equal to 77, including start and end symbols that delimited each word in the corpus.

3.2 Embedding Layer

The embedding layer projects each phonetic segment in the input into a continuous representation that is passed along to the recurrent layers. The embedding layer has 68 nodes: twice the number of phonological features in the feature representation that we chose (described in more detail in Section 4.1). Since the input layer uses a one-hot representation, this means that every phonetic segment in the inventory is represented as a vector of 68 weights between the corresponding input node and the embedding layer—*i.e.,* an *embedding*. These weights were initialized according to the procedure described in Section 4.1. The activation function for nodes in this layer was linear, with a bias term of 0.

3.3 Recurrent Layers

Each of the two recurrent layers of the network consisted of 512 LSTM nodes. The number of recurrent layers, as well as the number of nodes in each layer, were determined through extensive hyperparameter tuning (see Table A1 for details).

Each LSTM node receives input not only from the embedding layer, but also from its previous state. This allows the network to maintain a history, keeping track of the phones in the word up to the current point. Compared to simple recurrent neural networks, LSTMs have proven better at learning longer-distance dependencies, allowing them to represent more complex dependencies across non-adjacent timesteps (Hochreiter and Schmidhuber, 1997).

3.4 Output Layer

The output layer is a linear decoder layer as large as the segment inventory: 77 nodes. As in the input layer, each node corresponds to a particular phonetic segment. The output of the entire model, then, corresponds to a probability distribution over the next segment. This distribution is normalized using a softmax function, and the cross-entropy between this normalized distribution and the one-hot vector of the actual next segment indexes the accuracy of the model's prediction.

4 Experiment 1: Evaluating on a Held-Out Test Set

To investigate whether pre-specified distinctive features are helpful for acquiring phonotactic generalizations, we created two versions of a phonotactic learner: one that initially represents incoming phonetic segments as distinctive feature bundles (a *feature-aware* condition), and one that ini-

[3] All source code for models, training/validation/test sets, result files, and analysis scripts are included as supplementary material and freely available on GitHub.

tially represents phonetic segments as random vectors (a *feature-naive* condition). Our experimental manipulation occurs in the initialization of the weights between the input layer and the embedding layer; all other parameters were held constant. To compare these, we trained them on a identical subsets of the CELEX2 corpus (Baayen et al., 1995), and evaluated the likelihood that each model assigned to a non-overlapping test subset from the same corpus.

4.1 Method

Training Procedure

All models had the structure described in Section 3. Before training, the value of the weights between the input and embedding layers was determined in one of two ways, depending on the experimental condition to which the network was assigned:

1. **Feature-aware condition**: The weight vector of each phonetic segment was determined according to its distinctive feature specification, according to the scheme described later in this section (see "Distinctive Features"). Each weight was initialized as either -1, 0, or 1, depending on the phonetic segment's value for the feature in question.

2. **Feature-naive condition**: The weight vector of each phonetic segment was populated randomly from a distribution over the values -1, 0, and 1, with proportions identical to those found in the feature-aware condition.

All other weights were initialized from a uniform distribution between $\pm h^{-1}$, where h was the number of nodes in the subsequent layer.

All weights in the network were adjusted via backpropagation during the course of training. These included the weights between each layer, as well as the weights between successive states of the recurrent layers and those controlling each gate of each LSTM node. The error function used for this was cross-entropy loss, calculated over the 77 phonetic segment classes. Minimizing this cross-entropy loss is equivalent to maximizing log likelihood.

Each word in the training corpus was treated as a minibatch, with stored error backpropagated through the network once per word using stochastic gradient descent. Activations in each layer were automatically reset after each backpropagation to random values that were generated at the beginning of training.

Through hyperparameter tuning (detailed in Table A1), we settled on 1.0 as a suitable value for the initial learning rate, and annealed this by a factor of 0.25 every time there was no improvement on the validation set. The aforementioned hyperparameter tuning also led us to employ a dropout of 0.2, adjusting only 80% of the training weights per minibatch. Each model was trained for a total of 25 epochs (complete runs through the training corpus), after which the iteration of the model that assigned the highest log likelihood to the validation corpus was evaluated on the held-out test corpus, and the phonetic segment embeddings were stored for further analysis (see Section 6).

Twenty-five random initializations were trained in both the feature-aware and feature-naive conditions, for a total of 50 initializations. Within a condition, each initialization varied with respect to the initial weights except those between the input layer and the embedding layer.

Data

Corpus We used a randomly selected 50,000-lemma subset from the English part of the phonetically-transcribed CELEX2 database (Baayen et al., 1995) to train and test our model[4]. 30,000 of these lemma words were used to train the model, and the remaining 20,000 were randomly divided into validation and test sets of 10,000 lemmas each. Lemmas were used instead of inflected forms in order to minimize the number of shared stems across the three sets.

The only preprocessing steps applied to these data were the translation of each lemma from the DISC notation used in CELEX2 into IPA (with diphthongs split into separate phonetic segments, in order to increase comparability with Futrell et al., 2017) and the addition of start and end symbols around each word. No syllabification was added, because the models should infer the shape of syllables from the data alone, due to their ability to represent information across multiple timesteps.

Distinctive Features The precise distinctive feature structure we used to initialize the phonetic segment embeddings was based on Futrell et al. (2017)'s hierarchical feature dependency graphs,

[4] The CELEX2 corpus was also the basis of Futrell et al. (2017)'s data set.

in order to compare our model to this prior work. In these graphs, each node represents a feature, and certain features are only defined if their ancestor nodes have a certain value. For example, the "height" node is only defined if the manner of segment at hand is "vowel"; this is because the manner node is an ancestor of the height node.[5]

The first modification that we made to these feature dependency graphs is representing each multivalent feature as a binary one. This is because the values of several features do not lie along a straightforward unidimensional continuum. For instance, the "manner" node specifies the manner of a syllable, and has "trill" and "fricative" as two of its values. These manner classes are equivalent in terms of the size of the articulatory aperture: their ordering along a unidimensional continuum would be totally arbitrary. Instead, we split each possible value of a multivalent feature into a set of binary features, of which only one can be positive (1) at a given time; the rest must be negative (-1), if the feature is defined for the segment at hand.

In translating these dependency graphs into vectors, we represent each feature as a pair of dimensions in each phonetic segment vector. The first dimension in each pair expresses the value of the node: positive (+1), negative (-1), or unset (0). The second dimension in each pair denotes whether the node is set (1) or unset (-1), allowing for privative feature representation. This auxiliary dimension may seem redundant, but we include it because it is not the case that unset feature values are truly 'intermediate' between positive and negative ones, as a representation without the auxiliary dimension would suggest. We also add another two pairs of dimensions to represent start and end symbols.

Dependent Variable

We used log likelihood on the held-out test corpus of 10,000 lemmas to evaluate the quality of our models' phonotactic generalizations. The more accurate a model's representation of English phonotactics is, the higher the likelihood it should assign to extant English words that it has not seen.

4.2 Results

Performance of each model is plotted in Fig. 1. Using a Wilcoxon rank sum test with a continuity correction, we find that models in the feature-naive condition assigns a significantly higher log likelihood to the test corpus than those in the feature-aware condition ($W = 2.43 \times 10^{10}; p < .001$). On average, the feature-aware models assigned a log likelihood of -20.98 to the words in the test set, and the feature-naive models assigned an average log likelihood of -20.07. In other words, the feature-naive models assigned over twice the probability mass to the test set compared to the feature-aware models, in terms of raw (non-log) probability. The poorer performance of the feature-aware condition suggests that distinctive features need not be specified *a priori* of training, and that in fact they may bias the model toward suboptimal solutions.

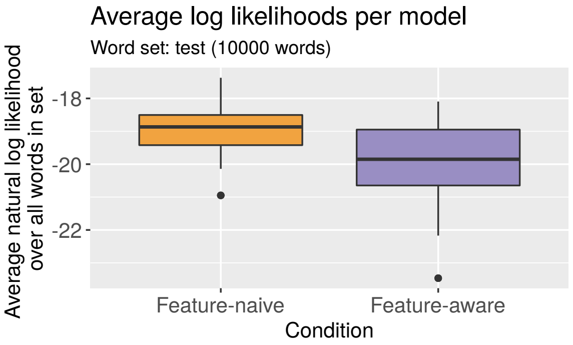

Figure 1: Box-plot of log likelihoods per model in each experimental condition. Each observation used to generate this plot ($N = 50$) is the average log likelihood assigned to each word in the test set, for a single model.

5 Experiment 2: Comparison to Human Judgments

In an effort to validate our models externally against evaluations that humans make, we ran another experiment correlating our models' log-likelihood ratings of non-words to human word-likeness judgments of the same non-words.

5.1 Method

Stimuli

Non-words were designed by Daland et al. (2011) to vary in the level of sonority sequencing principle violation, and as such their form was quite constrained: 96 stress-initial CCVCVC non-words, each starting with a consonant cluster that was either unattested (18 clusters), marginally attested (12 clusters), or frequently attested (18 clusters) as an onset in English. No non-word had more than one lexical neighbor, and non-words whose first

[5] These features are detailed further in Graff (2012), though some have been omitted since they are not distinctive in English.

or last 4 segments formed a existing word were excluded.[6]

Procedure

All human data for this experiment was collected by Daland et al. (2011). Forty-eight participants were recruited through Amazon Mechanical Turk; results were only retained from those reporting high (N = 2) or native (N = 36) English proficiency. Each participant performed a Likert wordlikeness rating task (1–6, where 6 was more wordlike) on all 96 stimuli, followed by a head-to-head comparison rating task in which participants were given two words and instructed to choose the non-word that seemed more like a typical English word. Each of the 4560 possible pairs was assigned to a single participant, and no participant saw any non-word more than twice during this task.

Daland et al. (2011) found that the comparison average of each non-word (proportion of comparison trials in which it was selected as better than its competitor) correlated with its average Likert rating across participants. However, the comparison average was more sensitive in differentiating non-words at the bottom of the Likert scale; therefore, we used the comparison average to evaluate our models.

Our models were the same feature-aware and feature-naive models from Experiment 1, trained on the same data. After training, we calculated the log-likelihood of each of the 96 non-word stimuli from Daland et al. (2011) for each of the 50 models from Experiment 1, and correlated these log-likelihoods to the human-derived comparison averages via the Spearman method.

5.2 Results

The correlations between the models' log-likelihood ratings and the human-derived comparison averages were moderate-to-strong, with Spearman's ρ ranging from 0.50 to 0.79, which is in the range of the best-performing models from Daland et al. (2011) that were trained on a comparable, but smaller, amount of unsyllabified data (20,000 vs 30,000 words). However, a Wilcoxon rank sum test on ρ yielded no significant difference between feature-naive and feature-aware models in this regard ($W = 282; p = 0.56$). This indicates that, although both feature-aware and feature-naive models can predict human judgments of non-words, the log-likelihoods assigned to this particular set of non-words do not distinguish the feature-aware from the feature-naive models.

6 Clustering of Learned Phone Embeddings

To examine the representations that are most helpful for characterizing word-level phonotactic generalizations, we performed a qualitative cluster analysis of the phonetic segment embeddings learned by the randomly-initialized model within each condition that assigned the highest average log likelihood to the test corpus.

First, we used agglomerative nesting to cluster the learned phonetic segment embeddings, which were grouped according to the Euclidean distance between them[7]. Position of each group was calculated in the 68-dimensional space via the unweighted pair-group average method (Sokal and Michener, 1958). The results are depicted in Fig. 2 for the feature-aware model and Fig. 3 for the feature-naive model. Comparing them, we see that the feature-aware model maintains manner-based distinctions even at late stages of the clustering. In contrast, these distinctions as not as clearly depicted in the feature-naive model, but it appears this model still encodes some phonetic information; namely: all stops are incorporated into the structure early, most vowels are incorporated into the structure after non-vowels, and several clusters contain only vowels of the same quality, collapsing over stress.

As clustering based on Euclidean distances is only a simplification over the non-linear transformations the network performs, this is a lower bound on the amount of structure the network can find. The feature-naive models' better performance on the test set suggests that these models may be encoding phonotactic-relevant knowledge in a more distributed representation that cannot be visualized thus—for example, a representation across several layers.

The phonetic information encoded by the models may be reflected in the heat map of feature embeddings, plotted in Figs. 4 and 5. To generate these, agglomerative clustering was performed in two dimensions: both on the phonetic seg-

[6] A full list of these words, as well as their wordlikeness scores, is downloadable from the first author's website.

[7] Clustering along Manhattan distance, as recommended by Aggarwal et al. (2001), yielded similar results.

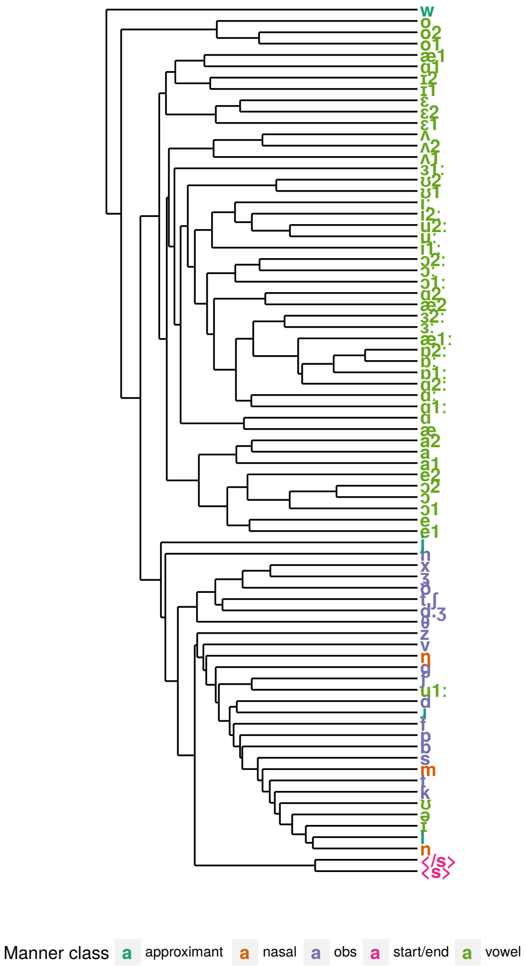

Figure 2: Dendrogram created using agglomerative clustering on trained embeddings from the feature-aware model that achieved the highest log likelihood on the test corpus. <s> and </s> signify start- and end-of-word symbols, respectively, and numbers after vowels indicate primary (1) and secondary (2) stress.

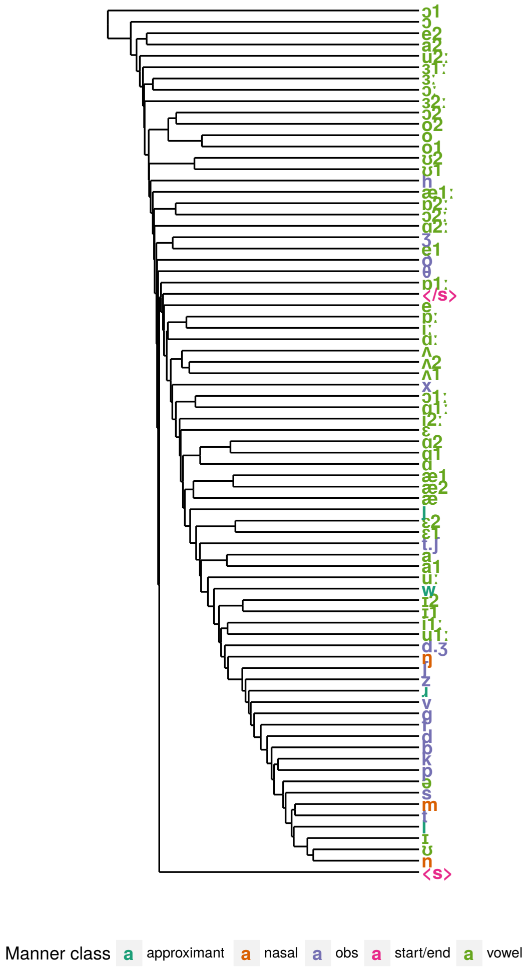

Figure 3: Dendrogram created using agglomerative clustering on trained embeddings from the feature-naive model that achieved the highest log likelihood on the test corpus.

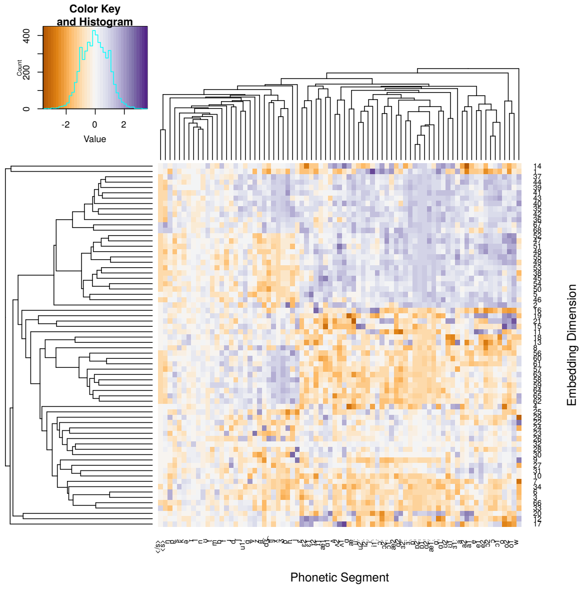

Figure 4: Heatmap of trained embeddings created from the feature-aware model that achieved the best performance on the test set. Clusterings along top axis are based on trained embeddings of each segment.

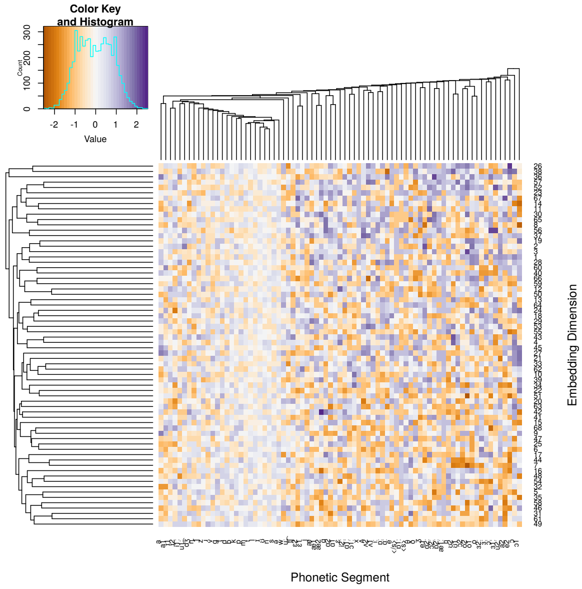

Figure 5: Heatmap of trained embeddings created from the feature-naive model that achieved the best performance on the test set.

ments and on the embedding dimensions. Colored patches of activations in these heat maps correspond to clusters of dimensions that all activate in response to certain phonetic segments—that is, clusters of dimensions that define a feature. Especially informative are patches with the same activation value below a cluster of phones: this means that the cluster is based on the feature encoded by those dimensions. For example, the two most well-defined final clusters formed by the feature-aware model are supported by multiple features, and Fig. 4 reflects this through wide horizontal bands that span the length of those clusters. Here, the vertical width of each band indexes the number of features that define the cluster.

The picture is much less clear for Fig. 5, which represents the embeddings learned in the feature-naive condition. The noisiness of the heat map indicates the clusters are not as distinct from each other: though every cluster is defined by at least one embedding dimension, these dimensions do not correlate in terms of their response to other phones outside the cluster. Instead of creating a straightforward clustering along embedding dimensions, the feature-naive model encodes any information that may be relevant to phonotactic probability in a more distributed representation.

7 Discussion

Returning to our initial questions, it seems prespecified phonological distinctive features are not required for phonotactic learning. All else being equal, representing phonetic segments as bundles of phonological distinctive features does not appear to aid in forming segment-level phonotactic generalizations, and, for this class of learning model, this specific distinctive feature set may even be detrimental. The fact that the feature-naive condition was able to encode phonotactic patterns indicates that all data required to represent these patterns as probabilities between phonetic segments is present in the sequence of segments itself; the learner need not rely on external information, such as distance between phones in acoustic space.

This is not to say that phonetic information is irrelevant to phonotactic learning. From examining the encodings that are learned during this process, we observe that the best models do encode some phonetic data.

This work is an example of how the initialization of even a single layer of a deep learning model can affect its ultimate performance on a held-out test set, a fact already demonstrated and discussed by, for instance, Sutskever et al. (2013). This effect was not observed in the models' correlations with human judgments, but this may be due to the limited number and form of non-words tested; with more statistical power, this measure may gain enough precision to distinguish the two conditions[8].

Finally, most of our models do assign a higher log likelihood to the test corpus than Futrell et al. (2017), which achieved a log likelihood of -21.73, suggesting that neural networks may be just as good at capturing phonotactic regularities as models that generate upcoming phonetic segments via stochastic memoization. However, our initial training set was much larger; when trained on only the 2,500 lemmas in Futrell et al.'s training set, our models yielded slightly lower log likelihoods than theirs (though we could not compare directly because their test set was inaccessible).

8 Implications

Per our results, phonological distinctive features do not appear to be mandatory for phonotactic acquisition. At the segment level, phonotactic patterns are learnable from distributional characteristics of each segment alone. This signals a need for revision of segmental phonotactic learning models that rely on a set of predetermined distinctive features—or at least stronger justification for the inclusion of any proposed distinctive feature set over another.

There are still a few additional tests that must be done before our conclusions can be generalized beyond these experiments. First, although the feature set that we used is typical of those used by other models of phonotactics, it is still possible that some other phonological feature set would result in better performance. Second, distinctive features may yet be helpful for models that train on much smaller datasets than ours, since they can provide hints to phonological structure that are not inferrable from such limited data.

Beyond distinctive features, some other, more detailed phonetic representation may yet prove helpful for phonotactic acquisition, if phonotactic expectations actually contain more detail about token-level variability, instead of the discrete segment-level representation assumed herein. Precise consequences for extant phonotactic learning models will depend whether this is the case; the determination is complicated by the fact that humans acquire both phonetic categories and phonotactic patterns simultaneously (Jusczyk et al., 1994; Werker and Tees, 1984).

One interesting avenue for future research is the multi-language case—i.e., training the model on a corpus in one language, and analyzing its performance on a corpus in a different language. This can help us make predictions about the types of pronunciation difficulties that speakers are likely to encounter in a second language, illuminating phonological effects of cross-linguistic transfer.

We must nonetheless be wary of using these results to make claims about human language acquisition. Human language is shaped by many other factors that are extraneous to our models, including articulatory restrictions, perceptual limitations, and constraints of cognitive economy. At the risk of overreaching, we must better specify these factors and their consequences before drawing further analogies.

9 Conclusion

Phonotactic acquisition can be accomplished without external, prior knowledge of distinctive features; indeed, according to our results, this knowledge may be a slight hindrance rather than a help. Though segment-level phonotactic inference may still benefit from access to a finer-grained phonetic specification of the speech stream, a predetermined encoding of this input in terms of distinctive features does not appear to be required for this purpose.

Acknowledgements

We thank Matthew Goldrick, members of the Northwestern Language & Computation Lab and Northwestern SoundLab, as well as the audience at MidPhon 2018 and anonymous reviewers for their helpful feedback. We also wish to thank the developers of the PyTorch Word-Level language modeling RNN example, which served as a starting point for our RNN code. This research was supported in part by NSF GRFP Award DGE-1842165 (Mirea) and NSF 1734217 (Bicknell).

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Jan Van den Bussche, and Victor Vianu, editors, *Database Theory — ICDT 2001*, volume

[8] This homogeneity may not have been an issue for Daland et al. (2011)'s comparison because the models tested therein had very diverse structures, and may have become sensitive to very different aspects of the training data as a result.

1973, pages 420–434. Springer Berlin Heidelberg, Berlin, Heidelberg.

Adam Albright. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1):9–41.

John R. Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

R H Baayen, R Piepenbrock, and L Gulikers. 1995. CELEX2.

Amélie Bernard. 2017. Novel phonotactic learning: Tracking syllable-position and co-occurrence constraints. *Journal of Memory and Language*, 96:138–154.

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

G. Nick Clements. 2009. The Role of Features in Phonological Inventories. In Eric Raimy and Charles E. Cairns, editors, *Contemporary Views on Architecture and Representations in Phonology*, pages 19–68. MIT Press.

Jeffry A. Coady and Richard N. Aslin. 2004. Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89(3):183–213.

Robert Daland, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology*, 28:197–234.

B. Elan Dresher. 2015. The arch not the stones: Universal feature theory without universal features. *Nordlyd*, 41(2):165–181.

Ewan Dunbar, Gabriel Synnaeve, and Emmanuel Dupoux. 2015. Quantitative Methods for Comparing Featural Representations. In *Proceedings of the International Congress of Phonetic Sciences*.

Stefan A. Frisch, Nathan R. Large, and David B. Pisoni. 2000. Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of memory and language*, 42(4):481–496.

Richard Futrell, Adam Albright, Peter Graff, and Timothy J. O'Donnell. 2017. A Generative Model of Phonotactics. *Transactions of the Association for Computational Linguistics*, 5(0):73–86.

Matthew Goldrick. 2004. Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language*, 51(4):586–603.

Peter Nepomuk Herwig Maria Graff. 2012. *Communicative Efficiency in the Lexicon*. Thesis, Massachusetts Institute of Technology.

Bruce Hayes and Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry*, 39(3):379–440.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Peter W. Jusczyk, Paul A. Luce, and Jan Charles-Luce. 1994. Infants′ Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, 33(5):630–645.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2741–2749. Association for the Advancement of Artificial Intelligence.

Jeff Mielke. 2008. *The Emergence of Distinctive Features*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford, New York.

Jeff Mielke. 2012. A phonetically based metric of sound similarity. *Lingua*, 122(2):145–163.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38(1):1409–1438.

Holly L Storkel, J Armbrüster, and Hogan, T P. 2006. Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language & Hearing Research*, 49(6):1175–1192.

Martin Sundermeyer, Hermann Ney, and Ralf Schlüter. 2015. From Feedforward to Recurrent LSTM Neural Networks for Language Modeling. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):517–529.

Ilya Sutskever, James Martens, and George Dahl. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of Machine Learning Research*, volume 28, page 9, Atlanta, Georgia, USA.

Michael S. Vitevitch, Paul A. Luce, Jan Charles-Luce, and David Kemmerer. 1997. Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Language and Speech*, 40(1):47–62.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

A Appendix

Hyperparameter Name	Description	Values Tested
`rand_reset`	whether activations in the model reset to a random state (True), or to zero (False) after each word	**True**, False
`lr`	initial learning rate	0.1, 0.5, **1.0**, 2.0
`anneal_factor`	amount by which to anneal learning rate, if no improvement found	0, **0.25**, 0.5, 1.0
`patience`	number of training epochs to wait for validation loss to improve before updating weights	**0**, 2, 4
`dropout`	proportion of weights to keep fixed	0, **0.2**, 0.5
`epochs`	number of epochs (complete passes through the data) to train for	**25**, 50, 100
`nlayers`	number of recurrent layers	1, **2**, 4
`nhid`	number of nodes in each recurrent layer	128, 256, **512**, 1250

Table A1: Particulars of hyperparameter testing. Hyperparameters were optimized for speed and likelihood assigned to the validation set. Optimal parameters for the validation set are bolded, and were used in the experiments reported here.