A Richer Model is Not Always More Accurate: Evaluating Phonotactic Knowledge with 8,400 Nonwords

Calvin Yang Kevin Tang
cyang2@ufl.edu tang.kevin@ufl.edu
University of Florida

Motivation: Wordlikeness judgements have been used as a primary tool to evaluate the nature of speakers' phonotactic knowledge. Gradient models of phonotactics are typically variants of a probabilistic language model computed over sounds at different levels of granularity, such as segments, and phonological features. A range of models have been proposed and shown to capture aspects of the phonological grammar. However, efforts in evaluating how existing models of phonotactics perform on wordlikeness judgements are limited in several ways. Firstly, they typically focus on specific phonotactic patterns and are evaluated over a small set of judgement data. For instance, Daland et al. (2011) evaluated different phonotactic models with 95 English nonwords for the sonority projection effects and, similarly, Gorman (2012) evaluated with 187 English monosyllabic nonwords for gross phonotactic violations. Secondly, studies typically evaluate a limited set of models/parameters. Thirdly, studies rarely always control for analogical learning effects. Finally, non-linear effects of phonotactic variables on wordlikeness are captured poorly with variable transformations such as logarithm. This paper overcomes these limitations by holistically evaluating a range of phonotactic models using a large nonword judgement database of 8,400 nonwords which are phonotactically diverse (with gross and minor phonotactic violations) to provide a high variance for determining the performance of phonotactic models, while modeling the effects of non-linearity and analogy.

Background: The classical N-gram model (Jurafsky & Martin 2009) is an ngram language model computed over segments. Variants of it tend to enrich the representation, for instance, the feature-based ngram model (Albright, 2009) captures the probability of a segment in a natural class. Besides the richness of the representation, some models use more complex architectures. For instance, the Hayes & Wilson (2008) Phonotactic Learner is a constraint-based learning model computed over features. While not commonly used in phonology, the Naive Discriminative Learning (NDL) model (Baayen et al., 2013) is proposed to be a psychologically plausible model of human learning. NDL is a two-layer wide network without immediate layers where association weights are learned from cues (features/segments) to outcomes (real words) using the Rescorla-Wagner learning rule (Rescorla, Wagner et al., 1972). The latent variables extracted from the association weights have shown to be predictive of lexical behavioral tasks (Milin et al., 2017). Recently, deep learning neural models such as Simple Recurrent Neural Network (sRNN) have shown potential for modeling phonotactics. Mayer & Nelson (2019) proposed two variants of sRNN by representing segments as embeddings conditioned on how they distribute among other segments or as features.

Research question: We ask whether a richer model is more accurate in terms of a) the level of linguistic representation and b) the model's architecture.

Methods: We conducted a series of model comparisons with models of two levels of representations (segments and features) and three architectures (N-gram, NDL and sRNN). The evaluation of the H&W model is currently under way. N-gram was computed over bigram-level information, NDL was trained on bigram-level cues to predict a word, and sRNN uses all preceding information of a segment to predict the segment. **Experiments:** Experiment 1 evaluates whether

feature-based models will outperform segment-based models by keeping the architecture constant: a) segmental vs. featural bigram, b) segmental vs. featural NDL, and c) segmental vs. featural sRNN. Experiment 2 evaluates whether sRNN outperforms simpler models such as NDL and N-gram. **Data:** A nonword judgement database of with 8,400 English nonwords (Needle et al., 2018) was used. The nonwords were normed using a rating of 'English-like-ness' on a 5-point Likert scale on average by 24 participants. The ratings were z-transformed and averaged across participants. The models were trained on the CMU English lexicon without stress and syllabification and the feature set defined in Hayes (2009). **Modeling:** Following the footsteps of Gorman (2012) and Harris, Neasom and Tang (2016), an analogical learning model which is distinct from phonotactic learning will serve as a baseline of wordlikeness and will allow for a conservative evaluation of phonotactic models by first taking the effect of analogical learning into account in a regression model. The Generalized Neighborhood Model (GNM) by Bailey and Hahn (2001) was chosen as our analogical model, since it is a rich model that compares each nonword with all the words in the lexicon weighted by their phonological distances. To capture potential non-linear effects these models have on wordlikeness judgement, Generalized Additive Models were used to predict wordlikeness ratings with phonotactic variables and the GNM variable as smooth terms and tensor product interactions between each variable and word length (Daland, 2015). Nested model comparisons were used to evaluate the importance of a phonotactic model. In Exp. 1, superset models with both the segmental and the featural phonotactic variables were fitted for each architecture. In Exp. 2, a superset model with the best phonotactic variables from each three architectures in Exp. 1 were fitted. Subset models were fitted by dropping one variable at a time. Changes in AIC and R^2 when a model's variable is dropped were used to measure variable importance (more changes = more important).

Results: **In Exp. 1**, segmental models performed better than featural models for N-gram (segmental: ΔAIC: 80, ΔR^2: 0.81%; featural: ΔAIC: 4, ΔR^2: 0.14%) and NDL (segmental: ΔAIC: 285, ΔR^2: 2.38%; featural: ΔAIC: 67, ΔR^2: 0.56%). The reverse is true for sRNN with the featural model being more important (featural: ΔAIC: 111, ΔR^2: 0.96%; segmental: ΔAIC: 57, ΔR^2: 0.53%). **In Exp. 2**, segmental NDL was found to be the best model of the best models in Exp. 1 (ΔAIC: 250, ΔR^2: 2.09%), followed by featural sRNN (ΔAIC: 186, ΔR^2: 1.49%) and segmental bigram (ΔAIC: 16, ΔR^2: 0.19%). Finally, orthogonal to the question of model complexity, GNM outperformed segmental NDL by three-fold (GNM: ΔAIC: 1116, ΔR^2: 9.63%; NDL: ΔAIC: 437, ΔR^2: 3.67%).

Conclusions: By holistically evaluating three types of phonotactic models, our findings suggest that featural information does not guarantee an improvement in model fit. In fact, the best phonotactic model was trained over segments. This is surprising because it is commonly assumed that speakers' ability to generalize over segments plays an important role in phonological learning and generalizations rely on featural information. NDL, a model of a wide learning network with a cognitively motivated learning rule, outperformed two cognitively unmotivated models (sRNN and N-gram). This suggests that phonotactic learning models should center on cognitive plausibility. Together, these findings suggest that a model with a richer representation or architecture is not always more accurate.

Key references: *Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology, 28*(2), 197-234. *Mayer, C., & Nelson, M. (2019) Phonotactic learning with neural language models. In *Proceedings of the Society for Computation in Linguistics*. *Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, H. (2017). Discrimination in lexical decision. *PloS one*.