Learning nonlocal phonotactics in Strictly Piecewise phonotactic model

Huteng Dai
Rutgers University, New Brunswick
huteng.dai@rutgers.edu

Phonotactic learning is a crucial aspect of phonological acquisition and has figured significantly in computational research in phonology (Prince & Tesar 2004). However, one persistent challenge for this line of research is inducing non-local co-occurrence patterns (Hayes & Wilson 2008). The current study develops a **probabilistic** phonotactic model based on the Strictly Piecewise class of subregular languages (Heinz 2010). The model successfully learns both segmental and featural representations, and correctly predicts the acceptabilities of the nonce forms in Quechua (Gouskova & Gallagher 2020; G & G henceforth).

Quechua: In Quechua, stop-ejective and stop-aspirate **subsequences**, e.g. *k…k', are ill-formed (G & G) (Stops includes plain voiceless stop, ejective, and aspirated stop):

- Stop-ejective: **kuk'u*, **k'uk'u*, **k^{h}uk'u*;
- Stop-aspirate: **kuk^{h}u*, **k'uk^{h}u*, **k^{h}uk^{h}u*.

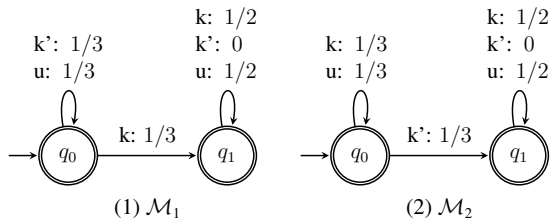
Hayes & Wilson (2008)'s baseline Maximum Entropy (MaxEnt) learner **locally** evaluates the bi-/trigram constraints. As distance increases, the search space grows so quickly that it becomes intractable; their learner cannot efficiently detect co-occurrence patterns over arbitrary distances. G & G offer a method for inducing tiers from placeholder trigrams (see also Jardine & McMullin 2017), however their learner is only shown to succeed on data in which the target phonotactics largely occur in local trigrams. In contrast, the current study induces non-local phonotactics by incorporating a Strictly Piecewise grammar from formal language-theoretic study into a probabilistic phonotactic model (Heinz 2010; Heinz & Rogers 2010).

SP phonotactic model: Strictly Piecewise (SP; or Precedence) grammar evaluates the **subsequences** of a string (Heinz 2010). A subsequence keeps track of the order, but not distance, between two elements in a string, e.g. k...k' in *kuk'u. A SP phonotactic model consists of a set of Probabilistic Deterministic Finite-state Automata (PDFAs) (Heinz & Rogers 2010). Each automaton checks if the symbol on the edge from state q_0 to q_1 is recognized. The transition probabilities are free parameters which are similar to **weighted constraints** in Harmonic Grammar (Legendre *et al.* 1990). The SP phonotactic model evaluates the **co-emission probability** $\mathsf{Coemit}(\sigma_i)$, which is the probability that all of the factored PDFAs emit a symbol σ_i at the same time (Shibata & Heinz 2019). The word likelihood is the product of $\mathsf{Coemit}(\sigma_i)$ of all segments in the word. The **feature-based** representation is also implemented by replacing the alphabet with a set of feature values $[\alpha F]$ (Heinz & Koirala 2010).

Figures (1, 2) show the SP phonotactic model banning *k…k' and *k'…k' with a simplified alphabet {k, k', u}. Figure (3) shows the derivation of *kuk'. $q_0 \xrightarrow[1/2]{\text{k}} q_1$ means "enter state q_1 from state q_0, emit symbol k and output probability $1/2$". The word likelihood of *kuk' is $0 = 1/3 \cdot 1/2 \cdot 0$. This model captures nonlocal phonotactics regardless of the amount of segments intervening between k and k', and always assigns 0 to k' after stops.

Learning: The learning problem of the SP phonotactic model is to estimate parameters i.e. transition probabilities so that the generated distribution maximally approaches the target distribution. As in Hayes & Wilson (2008), the parameters are optimized by minimizing the **negative log likelihood** (nll) of the learning data. The *Adam* algorithm (Kingma & Ba 2014) is applied to this optimization problem.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 401-402.
Held on-line February 14-19, 2021

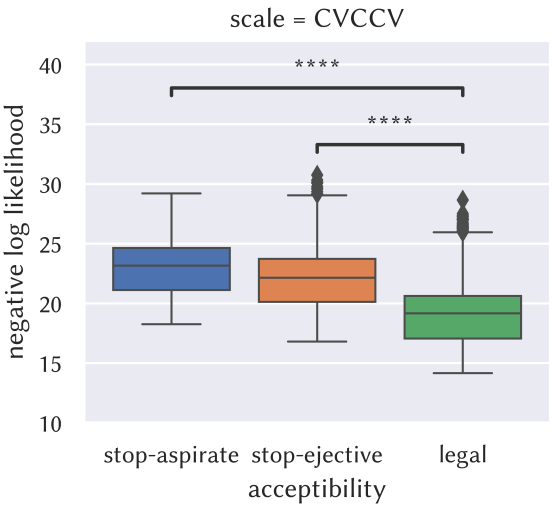


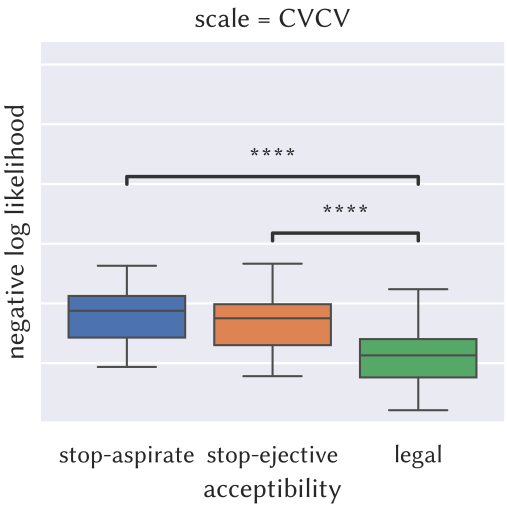
$$\mathcal{M}_1\colon\quad q_0 \xrightarrow[1/3]{\text{k}} q_1 \xrightarrow[1/2]{\text{u}} q_1 \xrightarrow[0]{\text{k'}} q_1$$

$$\mathcal{M}_2\colon\quad q_0 \xrightarrow[1/3]{\text{k}} q_0 \xrightarrow[1/3]{\text{u}} q_0 \xrightarrow[1/3]{\text{k'}} q_1$$

$$\textsf{Coemit}(\sigma_i)\colon \epsilon \xrightarrow[1/3]{\text{k}} \sigma_1 \xrightarrow[1/2]{\text{u}} \sigma_2 \xrightarrow[0]{\text{k'}} \sigma_3$$

(3) The derivation of *kuk'





The gradient is obtained by the AUTOGRAD package in PyTorch, which provides automatic differentiation for all operations in calculating nll. The SP phonotactic learner was applied to the Quechua dataset in G & G. The training data includes 10848 legal words. The testing data consists of 24352 nonce forms manually labelled as legal ($N = 18502$), stop-aspirate ($N = 3645$), and stop-ejective ($N = 2205$), and is further divided into CVCCV vs. CVCV based on syllabic structures as well as word length.

Primary result: Unlike the baseline grammar induced by the MaxEnt learner, the SP phonotactic model distinguish legal and illegal CVCCV nonce words in Quechua (G & G). The nlls of nonce words are clustered based on their acceptability, and the Mann–Whitney U test (McKnight & Najab 2010) is performed to test if the nll distributions of legal and illegal nonce words are signficantly different. In both segmental and featural model, SP phonotactic learner significantly distinguishes legal words from illegal stop-aspirate (Segmental: $p = 2.046 \cdot 10^{-185}$; Featural: $p = 2.113 \cdot 10^{-39}$) and stop-ejective (Segmental: $p = 2.945 \cdot 10^{-132}$; Featural: $p = 9.806 \cdot 10^{-37}$) pairs for either CVCCV or CVCV words. The clustering in segment-based model is illustrated in following boxplots. This result was considered only possible with tiers (Hayes & Wilson 2008; G & G), and the current study shows a promising alternative.

The parameters are interpretable as the probabilities of subsequences, e.g. $\textsf{Pr}([+\text{CG} \ldots +\text{CG}]) < \textsf{Pr}([+\text{CG} \ldots -\text{CG}])$ and $\textsf{Pr}([+\text{SG} \ldots +\text{SG}]) < \textsf{Pr}([+\text{SG} \ldots -\text{SG}])$ in the feature-based model. The result aligns with the generalized laryngeal phonotactics in Quechua.

Conclusion The current study rejects the claim that the formal language-theoretic (FLT) approach is incompatible with noisy corpus data (G & G) by implementing a probabilistic phonotactic model and learner. This learning relied on factored representations of the grammar made possible by the FLT study of these patterns (Heinz 2010; Heinz & Rogers 2010; Shibata & Heinz 2019). Furthermore, the SP phonotactic model excludes unattested *blocking effects* that are predicted by tier-based approaches (Heinz 2010).