Factoring syllabification into feature-based distributions describing phonotactic patterns

Cesar Koirala (koirala@udel.edu), University of Delaware, USA

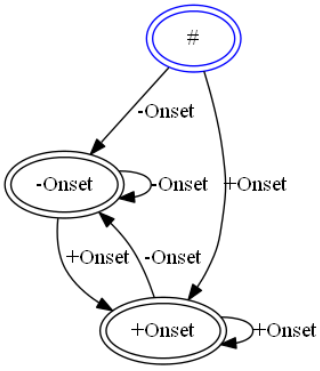
The role of syllables in phonotactics is a highly debated issue in theoretical phonology, and this debate has propagated itself into computational modeling. Recent work (Daland *et al.* 2011) has shown syllables to be useful, if not inevitable, in phonotactic modeling. Daland *et al.* (2011) show that feature-based models - Hayes & Wilson (2008) and Albright (2009) - unanimously benefit by using syllables when predicting sonority projection effects.

Motivated by aforementioned result, this paper attempts to incorporate syllabification using distributions of onsets and codas into a feature-based model - Heinz & Koirala (2010). An illustrative example comparing two versions of the Heinz & Koirala model (with and without syllables) is presented in order to identify the contributions syllables make.

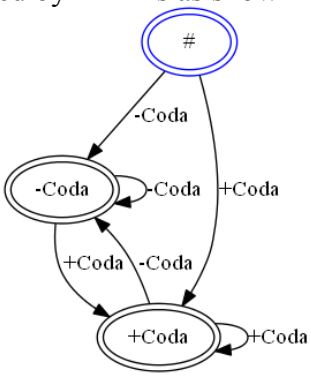
Heinz & Koirala (2010) is fundamentally different from feature-based models discussed in Daland *et al.* (2011). The idea is to identify the probability distribution of phonotactic patterns as a normalized product of simple distributions. These simple distributions could be distributions of individual features or distributions of a combination of features in a given sample. Probabilistic Deterministic Finite Automata (PDFA) can be used to represent such simple distributions, and the actual distribution is a particular product of these PDFAs. In the *baseline version* of the model, generalizations are based on the distribution of individual features.

The current work incorporates syllabification into the baseline version by utilizing the concept of syllable positions – onset and coda. For each segment, this positional information is encoded using a binary feature-like system. A consonant at the onset position of a syllable is encoded as <+ONSET, -CODA>, and a consonant at the coda position is encoded as < -ONSET, +CODA>. All the vowels are encoded as <-ONSET, -CODA>. This representation enables us to obtain two distinct probability distributions for the same consonant: one at the onset position and the other at the coda position. These distributions can be represented by PDFAs as shown in (1) and (2).

.



(1) Finite state machine for feature ONSET



(2) Finite state machine for feature CODA

The corpus is passed through these PDFAs. Then, the parse of the sample through these machines is counted and normalized in order to obtain the distributions of ONSETS and CODAs. Integrating these distributions into the product machine incorporates syllabification information. Here, syllabification is integrated into the model using two binary positional-features ONSET and CODA. This is just one out of multiple ways of integrating syllabification into this model,

and we choose this method in order to be consistent with binary feature representation in Heinz & Koirala (2010).

Size of the model: With 39 segments, an *n*-gram model which uses different symbols for segments in onset and coda positions has 78 states. As there are 79 associated probabilities at each state, the machine has 78*79=6162 parameters. In contrast, as the current model considers 20 features and 2 new positional-features (ONSET and CODA), there are only 2*(20+2) + 2^2*(20+2) +1= 133 parameters in a binary feature system (see Heinz & Koirala 2010 for validation). Hence, the parameters of this model can be estimated with much less training data.

Training the model: The model was trained on CELEX2 English lemma corpus with pronunciations taken from the CMU Pronouncing dictionary. As stress is not relevant, stress markings were removed. The training corpus consisted of 23,911 words. Training was done on two versions of the corpus. In one version, syllabification was specified (baseline version). In the second version, onset and coda positions were not distinguished in the training data (baseline_with_Syll version).

Illustrative Example: The following example with English consonant clusters illustrates the effect of incorporating syllabification information into the model. These *nasal-stop* clusters are prohibited at syllable onset position in English. The model with syllabification (4) distinguishes between these clusters at onset and coda positions by assigning different probabilities to them. The clusters have lower probabilities associated with them at the onset position compared to the coda position.

P (p | m)	75265 e-04
P (k | ŋ)	69442 e-04
P (t | n)	12966 e-03

(3) baseline version

P (p_onset | m_onset)	6.47855 e-05	P (p_coda | m_coda)	22911 e-04
P (k_onset | ŋ_onset)	6.41717 e-05	P (k_coda | ŋ_coda)	22630 e-04
P (t_onset | n_onset)	8.79082 e-05	P (t_coda | n_coda)	32692 e-04

(4) baseline_with_Syll version

Discussion and Conclusion: It was illustrated that syllabification can be integrated into a feature-based model as distributions of onsets and codas. Incorporation of syllabification provides important contextual information. In the illustrative example, incorporation of syllabification allowed the model to find different probability distributions for the same consonant cluster at different syllable positions with far fewer parameters than an *n*-gram model which uses different symbols for segments in onset and coda positions.

References

Albright, Adam (2009). Feature based generalization as a source of gradient acceptability. Phonology, Volume 26, Issue 01, pp 9-41.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis, & Ingrid Norrmann (2011). Explaining sonority projection effects. Phonology 28.pp 197–234.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. LI 39. pp 379–440.

Heinz, Jeffery & Cesar Koirala (2010). Maximum likelihood estimation of feature based Distributions. In proceedings of the Eleventh Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pp 28-37.