1) Abstraction-based Generalization

- Phonological rules can generalize to novel segments.
  - For example, when asked to add the possessive morpheme ‘-z’ to the name Bach, English speakers generalize a voicing assimilation process to the consonant /z/ (this is often called the Bach Test; Halle 1978).
- One explanation for this is that speakers represent the process using simple, abstract feature bundles:
  - For example: ‘[+voice]’ becomes [−voice] after [−voice].
  - I’ll call this hypothesis abstraction-based generalization.

- GMECCS (Moreton et al. 2017) is a maximum entropy (MaxEnt) phonotactic model that uses abstraction-based generalization.
  - GMECCS uses a gradient descent learning update, like most MaxEnt models.
  - The constraint set is every possible combination of relevant features (and their corresponding values): e.g. *[+voice], [+labial, +continuant], [+voice, -labial, -continuant], etc...
  - Crucially, there are simple, abstract constraints such as *[+voice], which cause generalization to novel sounds.

2) Similarity-based Generalization

- An alternative theory is that the generalization seen in phonology is due to “online calculations of similarity” (Cristia et al. 2013:279).
  - I’ll call this similarity-based generalization.
- Sim-Gen is a MaxEnt learner that differs from GMECCS in two ways:
  - Its constraint set only has single-segment constraints.
  - Each constraint’s weight updates leak onto featurally similar constraints.

3) Simulating the Bach Test

- To see if similarity-based generalization predicts the kind of behavior observed in the Bach Test, I trained GMECCS and Sim-Gen on an analogous phonotactic pattern.
  - Clusters made up of the segments {d, t, z, s, k, and g} all agreed in voicing.
  - Both models also had constraints that referred to [x] and [y], though these sounds were untested in the learning data.

4) Simulating Cristia et al. (2013)

- Cristia et al. (2013) taught participants an onset restriction (e.g. all onsets had to be [+voice]) and withheld three kinds of segments from training:
  - Within: within the legal onsets’ feature bundle. e.g. [b] if legal onsets were [+voice]
  - Near: phonetically similar to the legal onsets. e.g. [k] if legal onsets were [+voice]
  - Far: phonetically dissimilar to the legal onsets. e.g. [p] if legal onsets were [+voice]
- I simulated this experiment using GMECCS and Sim-Gen (η=0.05, 0=0.5). The figures below show the average probabilities each learner assigned to the various segment categories, averaged across all experimental conditions (probabilities are taken from each learner’s epoch that correlated most with the human data).

5) Discussion

- While both the similarity-based and abstraction-based learners were able to model the Bach Test, only the similarity-based learner was able to predict the results of Cristia et al. (2013).
- Future work:
  - Are there better similarity metrics? So far, I’ve tried similarity based on natural class membership (Frisch et al. 2004), articulatory phonetics (Mielke 2011), and lexical context (Mikolov et al. 2013) to no avail.
  - Sim-Gen captures the same complexity bias effects that Moreton et al (2017) found with GMECCS. Are there any biases that Sim-Gen might predict that other MaxEnt models don’t? 

References