A Probabilistic Model of Phonological Relationships

Kathleen Currie Hall
City University of New York – College of Staten Island & The Graduate Center
kathleen.hall@csi.cuny.edu

This paper presents a model of phonological relationships, the Probabilistic Phonological Relationship Model (PPRM), that precisely quantifies the degree to which two phonological units are predictably distributed in a language. Although it is widely accepted that (1) the ability to define phonological relationships such as contrast and allophony is crucial to the determination of phonological patterns in language (see, e.g., Goldsmith 1998) and (2) the notion of predictability of distribution is one of the key tools phonologists should use in determining phonological relationships (see, e.g., Steriade 2007), there are a large number of cases in the descriptive phonological literature that are problematic for the usual classifications of units as being either "predictably distributed" (allophonic) or "unpredictably distributed" (contrastive). To take one example, there is a long-standing debate about the status of the vowels [ai] and [ʌi] in Canadian English: Should they be considered allophonic because they are predictably distributed in *most* environments, or should they be considered contrastive because they are unpredictably distributed before flap in (near) minimal pairs such as *idol* [aiɾl̩] vs. *title* [tʌiɾl̩]?

The PPRM solves problems such as these by building on insights from probability and information theory to allow researchers to calculate intermediate degrees of predictability of distribution. The degree of predictability is quantified with the measurement of *entropy*, or the uncertainty of choice between the two sounds (e.g., Shannon & Weaver 1949). The three components of the model are given in (1-3). (1) shows the measure of bias toward *X* if the options are *X* and *Y*, in a given environment, *e*, based on the number of occurrences, *N*, of each; (2) shows the calculation of entropy, *H,* between *X* and *Y* in *e*; and (3) shows the calculation of entropy across environments given the probability, *p(e)*, of each environment. Note that because there are only two elements in question, entropy will range between 0 and 1 in this model.

(1) Bias (toward X): $p(X;\ (X \text{ or } Y)|e) = p(X_e) = N_{X/e} \ / \ (N_{X/e} + N_{Y/e})$
(2) Environment-Specific Entropy: $H(e) = -1 * \ (p(X_e) \log_2 p(X_e) + p(Y_e) \log_2 p(Y_e))$
(3) Weighted Average Entropy Across Environments: $H = \sum (H(e) * p(e))$

The calculations of probability and uncertainty in the model are made over corpora of language data, allowing the phonological relationships determined by the model to accurately reflect the distributions of units in the speech of language users. Using corpora also allows frequency information, long known to play a role in phonological processing, to be incorporated into the model.

Using the PPRM for Canadian English, it will be shown that the relationship between [ai] and [ʌi] is one of partial predictability. Specifically, [ai] and [ʌi] have a relatively high uncertainty of ~0.95 in the pre-flap environment, but the impact of this environment is relatively small, so that the overall uncertainty of the pair is very low (~0.01). As will be shown, this intermediate degree of predictability can simultaneously account for the facts that naïve Canadian English speakers can generalize the distribution of [ai] and [ʌi] to novel environments (e.g., Boersma & Pater 2007; Idsardi, 2006), a hallmark of allophony, but that the distribution of [ai] and [ʌi] is becoming less predictable and therefore apparently contrastive (Hall 2005). The PPRM can be applied to a large number of similar situations, thus providing a new and powerful tool for accurately describing relationships that hold between phonological units in a language.