Explaining non-native consonant cluster processing:
Bayesian combination of phonetic likelihoods and feature-based phonotactic priors

1. A Bayesian model of cluster processing

Native speakers of English and other languages produce systematic error patterns on a variety of tasks that involve the processing of non-native consonant clusters (e.g., Scholes 1966; Hallé et al. 1998; Pitt 1998; Dupoux et al. 1999; Moreton 2002; Davidson et al. 2004; Davidson 2006; Berent et al. 2007, 2008, 2009). For example, English speakers make many errors in attempting to produce initial [vd]; they also make errors, but fewer of them, on equally unattested initial [fn]. Previous research on such patterns has focused on comparing analyses that each rely upon a single factor: whole-segment frequency or transitional probability; phonetic properties of the constituent consonants and the transition between them; or selected phonological principles such as the Obligatory Contour Principle (OCP) on place of articulation and the sonority sequencing principle (SSP) (e.g., Moreton 2002; Davidson 2006; Berent et al. 2007; Peperkamp 2007). In this paper, we develop a formally explicit, multi-factor model in which phonetic properties of the cluster transition are combined with feature-based phonotactic generalizations according to Bayes' theorem. We demonstrate that the model accounts for a body of English speech production data better than several alternatives that differ with respect to the nature of the phonetic and phonological factors involved or the method by which they are combined. The proposed model is stated in (1a) and a linking hypothesis that relates it to the production data is stated in (1b).

(1) Bayesian model of cluster processing [where 'S' represents the stimulus]
   a. $\Pr([\#C_1C_2] \mid S) \propto \Pr(S \mid [\#C_1C_2]) \times \Pr([\#C_1C_2])$ [i.e., posterior ∝ likelihood × prior]
   b. Production error rate on $[\#C_1C_2]$ given S is directly proportional to the posterior $\Pr([\#C_1C_2] \mid S)$.

2. Specification of the phonetic likelihood

When S is an acoustic stimulus, the **likelihood** $\Pr(S \mid [\#C_1C_2])$ reflects the auditory compatibility between S and an intended phonological sequence in which initial C_1 is followed by C_2 with no intervening vowel and close transition (i.e., with a high degree of overlap between C_1 and C_2, in conformity with English onset gestural timing patterns; e.g, Byrd 1995). This compatibility will be lower — and the stimulus will be more consistent with alternative phonological sequences such as $[\#C_1{}^{ə}C_2]$ (excrescent schwa) or $[\#C_1əC_2]$ (epenthetic schwa) — to the extent that the C_1-C_2 transition in the stimulus contains properties indicative of open transition or the presence of a (short) vocoid. For the purposes of this paper, we focus on three properties of the transition that could negatively impact the likelihood $\Pr(S \mid [\#C_1C_2])$ relative to the likelihoods for other interpretations of S: (a) the place of C_1 is anterior to that of C_2, increasing the probability of an audible release of C_1 even under relatively close transition (Byrd 1992; Zsiga 1994); (b) C_1 or C_2 has formant structure, which could be misattributed to a vocoid between the two consonants; (c) C_1 or C_2 are voiced at the transition, which could also be misattributed to a transitional vocoid.

In order to derive predictions from (1a) under the linking assumption (1b), it is necessary to define a set of forms against which $[\#C_1C_2]$ competes for posterior probability given stimulus S. Here we adopt the idealization that the one alternative is $[\#C_1əC_2]$. The constant of proportionality in (1a) then becomes:

$$\Pr(S \mid [\#C_1C_2]) \times \Pr([\#C_1C_2]) + \Pr(S \mid [\#C_1əC_2]) \times \Pr([\#C_1əC_2])$$

This idealization is consistent with our attention to phonetic aspects of the transition that are most vocoid-like, and with the finding that the majority (though by no means all) of the production errors in the experiment described below involve some degree of epenthesis. However, we emphasize that our model is fully general and anticipate extending it to other types of production error in our talk presentation.

3. Specification of the phonotactic prior

The **prior** distribution of our model is determined by a phonotactic grammar. Concretely, we employ a revised version of the Maximum Entropy (MaxEnt) framework for phonotactics and phonotactic learning of Hayes & Wilson (2008). In this framework, the log probability of a phonological form is proportional to the negative sum of its weighted constraint violations. Constraints are stated in terms of features and other representational units of phonological theory and weighted according to the principle of maximum likelihood, which inherently favors more restrictive grammars. Because we are modeling the adult state of

the phonotactic grammar rather than its development, the constraints and their weights are learned from a lexicon of word types; however, the framework itself is compatible with learning from entire utterances, raising the possibility of extending our work to (errorful) child speech production. The present version of the MaxEnt framework diverges from that of Hayes & Wilson primarily in the mechanism of constraint induction: instead of the complicated search heuristics and O/E measure of the earlier work, we use a simpler and more principled algorithm in which constraints are selected based on the **gain** (Della Pietra et al. 2007) — or increase in log likelihood of the data — that they are estimated to produce. We propose a novel variant of gain-based selection in which constraints that are projected to have large weights (typically, unviolated constraints), and symbolically simpler constraints are preferred. These analytic biases can be formally interpreted as another Bayesian prior: a prior over phonotactic grammars.

The learning data and features for our simulations were similar to those of previous work. The feature set included distinctions for manner ([sonorant], [nasal], stricture features), place of articulation ([labial], [coronal], [dorsal], [anterior], [distributed]), and laryngeal and other features. Because the model contains no preference for constraints against OCP-place or SSP violations, well-formedness distinctions related to these principles must be induced from the data (together with the analytic biases outlined above).

4. Production data and modeling results

The Bayesian model of cluster processing was tested against the production data of Davidson (2006, Experiment 1). Davidson's experiment required native English participants (N=20) to produce four instances each of 24 non-native initial consonant clusters: [fk fm fn fp fs ft sf sk sm sn sp st vb vd vg vm vn vz zb zd zg zm zn zv]. The clusters were embedded in 96 $C_1C_2VC_3V$ nonwords, recorded by a Czech speaker, and presented to the participants auditorily. Participants' responses were recorded and judged incorrect if phonetic analysis revealed an open transition between C_1C_2 that included formant structure. Here we report our modeling results for the collapsed data (i.e., the error rate for each cluster collapsed across participants and stimulus items); a fuller presentation of the data, including mixed-effect modeling of participant and item variance, will be provided in the talk.

In applying (1) to the production data set, the priors $Pr([\#C_1C_2])$ and $Pr([\#C_1əC_2])$ were determined by the induced phonotactic grammar and, for mathematical convenience, the likelihood $Pr(S \mid [\#C_1C_2])$ was set to a constant value of 1.0. The likelihood $Pr(S \mid [\#C_1əC_2])$ was determined by an exponential function of four stimulus-based parameters: an intercept term $\beta_0 < 0$, which expresses a general incompatibility between the cluster stimuli and $[\#C_1əC_2]$, and β(front-to-back), β(formants), and β(voiced), each of which is constrained to be greater than 0 and active only when the C_1-C_2 transition has the corresponding property. The latter three factors express the greater acoustic compatibility between S and $[\#C_1əC_2]$ when there is a more robust release, formants, or voicing in the transition. The β parameters were fit by constrained maximum likelihood to the production data. The resulting model achieves a very good match (logLik = 89.0; Kendall's $\tau = 0.765$) to the error probability distribution. Likelihood-ratio tests between the full model and submodels lacking the phonotactic prior or one of the acoustic factors establish that all of the factors except voicing contributed significantly to the result. These findings support our main claim that non-native cluster processing reflects a mathematically justified interplay between phonetic interpretation and stochastic phonotactic knowledge. Several other models with non-featural phonotactic grammars and alternative constraint induction methods (e.g., those of Hayes & Wilson 2008 and Boersma & Pater 2008), when evaluated in the same way, were found to give substantially worse results.

Selected references

Byrd, D. 1992. Perception of assimilation in consonant clusters: a gestural model. *Phonetica* 49:1-24.
Berent, I., D. Steriade, T. Lennertz, and V. Vaknin. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104:591-630.
Davidson, L. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34:104-137.
Hayes, B. and C. Wilson. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3:379-440.