Bayesian modeling of cross-language speech processing

Colin Wilson (Johns Hopkins University)
Lisa Davidson (New York University)

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, Berent et al. 2007, Berent et al. 2008, Davidson 2010). 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 sonority sequencing principle (e.g., Moreton 2002, Davidson 2006, Berent et al. 2007). In this research, 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 alternatives which rely solely on phonological knowledge as the primary source of explanation.

The data that we aim to explain come from Davidson (2010), who required native English speakers to produce words with initial consonant clusters that are not possible in English (e.g. [bdagu], [tkale], [ftake], [zmago], etc). The stimuli were naturally produced by a native Russian speaker. Results showed that there were different rates of accuracy for different clusters. When speakers produced an error, it was most often insertion of vocalic material between the consonants, but other errors also occurred, such as prothesis or C1 deletion. However, a closer analysis of performance on individual tokens showed a surprising finding, illustrated by the data in Table 1: stimuli that begin with exactly the same phonological sequence (e.g., [zm] or [zn]) – but that differed in their fine phonetic details – were systematically produced with different modifications. In the case of [z]-nasal clusters shown in Table 1, the most relevant acoustic phonetic property is prevoicing: a short period of initial voicing that begins before frication (see Figure 1). Stimuli that contain prevoicing (e.g., [zmagu]) were produced with significant rates of prothesis; stimuli that do not contain prevoicing (e.g., [zmafo]) were not subject to prothesis, and most often produced correctly. Further analysis of both the stimuli and participants' responses reveals that different acoustic phonetic details are relevant for predicting modifications of other cluster types. For example, stop release duration and voicing is directly correlated with the rate of insertion modification, and the relativized burst/release amplitude of stops is inversely correlated with the rate of C1 deletion.

We show that neither phonological knowledge such as sonority sequencing nor lexically-based segmental information can adequately explain the speakers' behavior on the Russian stimuli. Instead, we argue that a successful account requires the integration of quantitative information about particular stimulus items, the perceptual likelihood, with a gradient measure of phonotactic well-formedness, the prior, as prescribed by Bayes' Theorem. We demonstrate that by modifying the Maximum Entropy (MaxEnt) modeling framework developed in Hayes and Wilson (Hayes and Wilson 2008), we can provide a good fit to the more detailed breakdown of the production data reported in Davidson (2010).

stimulus item	Correct	Prothesis	Epenthesis	C1 deletion
zmafo	.75	**.13**	.13	.00
zmagu	.36	**.64**	.00	.00
znafe	.78	**.00**	.22	.00
znagi	.45	**.55**	.00	.00

Table 1. Proportion of production responses to zN stimulus items

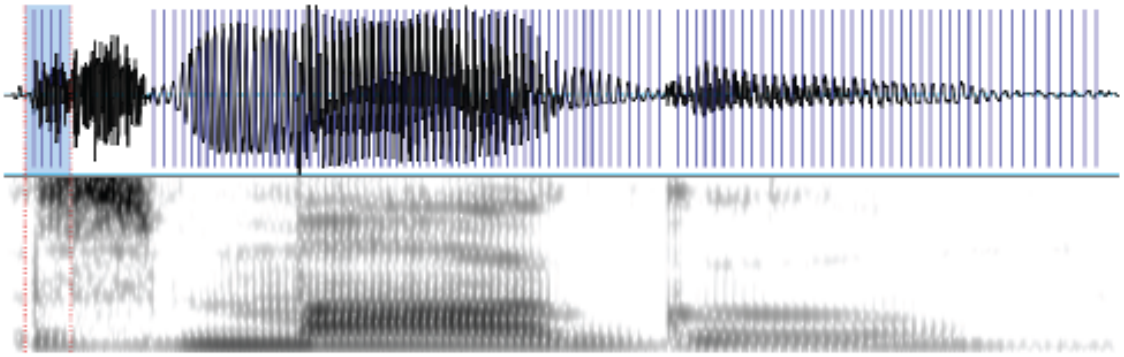

Figure 1. Spectrogram and waveform of the stimulus [zmagu]. Pitch pulses and formant structure indicate a period of prevoicing overlapping with frication.

References

Berent, I., T. Lennertz, J. Jun, M. Moreno and P. Smolensky (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences* **105**, 5321-5325.

Berent, I., D. Steriade, T. Lennertz and V. Vaknin (2007). 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* **34:1**, 104-137.

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* **38:2**, 272-288.

Dupoux, E., K. Kakehi, Y. Hirose, C. Pallier and J. Mehler (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* **25**, 1568-1578.

Hallé, P., J. Segui, U. Frauenfelder and C. Meunier (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance* **24:2**, 592-608.

Hayes, B. and C. Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* **39**, 379-440.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition* **84**, 55-71.

Pitt, M. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics* **60:6**, 941-951.

Scholes, R. (1966). *Phonotactic Grammaticality*. The Hague: Mouton.