Modeling the Perception of Vowel Nasalization

Yongqing Ye yeyongqi@msu.edu
Karthik Durvasula karthikd@msu.edu
Michigan State University

This paper formally models the time course of the perception of vowel nasalization within a Bayesian framework (Feldman 2009; Knill and Richards 1996; Norris and McQueen 2008, inter alia). It explores whether listeners (1) update inferences incrementally or rely solely on the immediate acoustics, (2) use time step-sensitive models of abstract categories, and (3) employ underspecified representations. Findings suggest that categorical, underspecified representations are sufficient to account for the patterns in listeners' perception during the time course of vowel nasalization, and that listeners engage in Bayesian inference where their decisions are continuously updated based on previous information.

Method: Forty-three native American English speakers participated in a production experiment, reading wordlists with multiple repetitions of 24 target CVC and CVN words, plus fillers. The same participants also completed a forced-choice perception task, identifying the final sound of end-truncated CVC and CVN nonce words at eight gates. Recordings from the production experiment were used to train acoustic models for each vowel category (CVC for oral vowels, CVN for nasalized vowels, and both together to model underspecified vowels), using Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein 1980) to generate likelihood distributions.

To simulate incremental perception, we combine two acoustic models with two Bayesian perception models: a time-sensitive acoustic model, featuring eight time-normalized multidimensional distributions per vowel category, and a non-time-sensitive acoustic model, using a single distribution per vowel category. Test items were mapped to these trained likelihood distributions, generating the probability of each item belonging to the corresponding vowel category. Each acoustic model was paired with either a non-updating prior Bayesian model, which used unchanging, equiprobable priors, or an updating prior Bayesian model, which uses the posterior at each time step/gate as the prior for the next time step/gate. The performance of these perception models was evaluated impressionistically against the results of the perception experiment.

Results: Models with dynamically updating priors, adjusting at each step based on prior computation outcomes, better reflected actual listener behavior than the non-updating models. This suggests that listeners engage in a form of Bayesian inference, where decisions are continuously influenced not only by the likelihood of the input given the known acoustic distributions but also by an evolving prior that incorporates newly accumulated probabilistic information at each step. Additionally, non-time-sensitive acoustic models yielded equally strong results, implying that listeners can sustain perceptual accuracy without depending on fine-grained temporal details, and that categorical representations may suffice to explain perception patterns without necessitating exemplars (Pierrehumbert et al. 2002; Pierrehumbert 2001, 2016) or claiming that listeners are attuned to nuanced phonetic details such as varying degrees of coarticulation (Fowler 1981, 1984, amongst others). Lastly, models incorporating underspecified representations yielded results comparable to those of fully specified ones, indicating that listeners may rely on underspecified representations effectively. Formal and explicit modeling of both the acoustic/auditory models used in perception and the perceptual computation over such acoustic models allows us to precisely define and articulate our underlying assumptions, decisions, and predictions when formulating and evaluating perceptual theories.

References

Davis, Steven and Paul Mermelstein (1980). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE transactions on acoustics, speech, and signal processing* 28.4, pp. 357–366.

Feldman, Jacob (2009). “Bayes and the simplicity principle in perception.” In: *Psychological review* 116.4, p. 875.

Fowler, Carol A (1981). “Production and perception of coarticulation among stressed and unstressed vowels”. In: *Journal of Speech, Language, and Hearing Research* 24.1, pp. 127–139.

Fowler, Carol A (1984). “Segmentation of coarticulated speech in perception”. In: *Perception & Psychophysics* 36.4, pp. 359–368.

Knill, David C and Whitman Richards (1996). *Perception as Bayesian inference*. Cambridge University Press.

Norris, Dennis and James M McQueen (2008). “Shortlist B: a Bayesian model of continuous speech recognition.” In: *Psychological review* 115.2, p. 357.

Pierrehumbert, Janet, Carlos Gussenhoven, and Natasha Warner (2002). “Word-specific phonetics”. In: *Laboratory phonology* 7.

Pierrehumbert, Janet B (2001). “Exemplar dynamics: Word frequency, lenition and contrast”. In: *Typological studies in language* 45, pp. 137–158.

Pierrehumbert, Janet B (2016). “Phonological representation: Beyond abstract versus episodic”. In: *Annual Review of Linguistics* 2, pp. 33–52.