A probabilistic model of phonetic cue restructuring

James P. Kirby (jkirby@uchicago.edu)
University of Chicago, Department of Linguistics,
1010 E. 59th St., Chicago, IL 60622 USA

Keywords: Phonetic change; speech perception; agent-based modeling; categorization; mixture models

Introduction

Research demonstrating that both infants and adults track statistical distributions of acoustic-phonetic cues and use this information when making phonetic category judgements (Maye, Werker, & Gerken, 2002; Clayards, Tanenhaus, Aslin, & Jacobs, 2008) has led to interest in computational models of phonetic category acquisition, which can shed light on the requirements and limitations of statistical learning. The results of a number of studies (Vallabha, McClelland, Pons, Werker, & Amano, 2007; Toscano & McMurray, 2008) have yielded encouraging results, indicating that Gaussian mixture (GMs) may be an appropriate means of representing phonetic category structure. However, these structures are not static; they can and do change over a speaker's lifetime, albeit in ways which are not yet fully understood. This work builds on previous research by embedding the GM approach in an agent-based framework to explore the ways in which phonetic category structure changes over time.

Sound change as phonetic cue restructuring

Speech sound categories (consonants and vowels) are not monolithic entities, but are instead signaled by a multitude of acoustic dimensions, called *cues*. Lisker (1978) cites 16 acoustic dimensions relevant for the perceptual distinction between voiced (e.g., [b]) and voiceless (e.g., [p]) obstruents in word-medial position in English, including duration of the preceding vowel, fundamental frequency (f_0) contour, and timing of voice onset (VOT). While many cues are truly independent, others, such as VOT and f_0 contour, are redundant: vowels following voiced obstruents have lower f_0 than vowels following voiceless obstruents; in addition, some cues contribute more information to the identity of a contrast than others. Accurate categorization of an utterance involves weighting these of these cues, a task which finds a natural analog in density estimation (Ashby & Alfonso-Reese, 1995) and closely related 'ideal observer' models of speech perception as optimal Bayesian inference (Clayards, 2008; Feldman, Griffiths, & Morgan, 2009).

The distribution of cues to a speech sound category are not static, however, and may shift and change over time. An oft-cited example is the idea that lexical tone – the use of pitch to distinguish between words, familiar from languages such as Mandarin Chinese or Thai – finds its origins in consonantally-induced pitch perturbations (Hombert, Ohala, & Ewan, 1979). On this account, the physiologically-based, consonantally-induced differences in vowel f_0 first become part of a perceptual cue distinguishing two types of consonants. If f_0 comes under speaker control, it may then be used to actively to enhance the perception of this contrast. More generally, when the primary cue to a contrast becomes uninformative, the contrast may still be maintained through increased attention to a secondary cue, a process termed *phonologization* (Hyman, 1976).

Table 1: Phonologization of f_0 in Seoul Korean.

manner		*1960s*	*2000s*	*gloss*
fortis	뿔	[ppul]	[púl]	'horn'
lenis	불	[pul]	[p^hùl]	'fire'
aspirated	풀	[p^hul]	[p^húl]	'grass'

Empirical support for such an account may be found in the phonologization of f_0 currently taking place in Seoul Korean (Kang & Guion, 2008). In this language, a three-way contrast between fortis, lenis, and aspirated word-initial voiceless obstruents once distinguished chiefly by differences in VOT is now distinguished chiefly by differences in f_0. As shown in Table 1, fortis and aspirated stops are both produced with high f_0, but distinguished along the VOT dimension, whereas lenis stops are distinguished from aspirated by low f_0.

As it happens, VOT and f_0 are not the only cues relevant for the perception of word-initial obstruents in Seoul Korean: a number of studies (reviewed in Kang and Guion (2008)) have shown that other acoustic characteristics, such as length of the following vowel and spectral tilt at vowel onset, are also important cues to obstruent category. If phonetic categories are signaled by a multiplicity of cues, however, it is not immediately obvious why should f_0, and not some other cue, should have been phonologized, nor why this change took place in Korean, but not in other languages which displays a similar redundancy between VOT and f_0, such as English.

In this work, I propose that this type of phonetic category restructuring is the result of an adaptive strategy of cue enhancement designed to ensure robust communication in noise. Speakers enhance phonetic cue dimensions probabilistically, in proportion to their contribution to the successful perception and categorization of a phonetic contrast, based on the *informativeness* of a cue and the *precision* with which the contrast may be recovered. This predicts that phonetic cue restructuring will result from the loss of contrast precision due to noise or external bias, with the degree of enhancement proportional to the loss of precision.

Modeling phonetic cue restructuring

A series of agent-based simulations were conducted to better understand the effects of probabilistic enhancement on cue weights. Five cue dimensions known to be relevant for the perception of Korean stops (VOT, vowel length, closure duration, spectral tilt, and f_0) were represented as a set D of three 5-dimensional GMs, corresponding to the three word-initial obstruent categories. Both the initial and target parameters of each GM were estimated from data in the apparent time study of Kang and Guion (2008), represented as an exemplar list $E_k = \{e_1^k, \ldots, e_n^k\}$, e_i^k a 5-dimensional column vector of cue values plus a category label k and a decay weight τ.

Agent-based simulations

The simulations reported here consist of simple 'telephone' conversations in which two agents alternate between producing and categorizing utterances. At each iteration, the speaker agent selects a phonetic category target k, computes maximum likelihood estimates of the parameters $\mu_{d|k}, \sigma_{d|k}$ for all $d \in D$ based on E_k, and samples from each conditional density $x_d \sim \mathcal{N}(d|k; \mu_d, \sigma_d)$ to generate an utterance vector $\mathbf{x}$. The agent then enhances cue dimension d of $\mathbf{x}$ with some probability, proportional to both (i) the cue's *weight* (based on normalized d') and (ii) the current contrast *precision*, defined as the error rate of a naive Bayes classifier. Finally, $\mathbf{x}$ may be further modified by a transmission bias term λ, used to implement systematic biases such as articulatory drift.

The utterance $\mathbf{x}$ is then presented to the listener agent for classification. The listener agent assigns a category label k with probability $P(k|x_1, \ldots, x_D)$, where the posterior probability of each category k is calculated as

$$P(k|x_1, \ldots, x_D) = \frac{p(x_1|k)p(x_2|k), \ldots, p(x_D|k)p(k)}{\sum_{i=1}^{K} p(x_1|k_i)p(x_2|k_i), \ldots, p(x_D|k_i)p(k_i)}. \quad (1)$$

After classification, the agent adds $\mathbf{x}$ to the top of the appropriate exemplar list E_k, re-computes decay weights, and deletes exemplars with sufficiently low τ (to simulate memory decay). In the next iteration, when the listener agent becomes the speaker, the contribution of this newly categorized exemplar will be reflected in production when the agent computes new maximum likelihood parameter estimates.

Results

Simulations of up to 50,000 iterations were conducted with and without enhancement and for various settings of the bias term λ. Neither the proposed probabilistic enhancement strategy nor systematic bias alone were sufficient to induce a shift in cue weights that resembled the empirical target distributions, but simultaneous application of both gave a close approximation of the attested distributions and cue weights, as measured by the Kullback-Leibler divergence between the simulated results and the empirical targets. A second series of simulations in which the weights of secondary cues to the contrast were equalized at initialization, systematic bias in the production of VOT (the primary cue) led to either partial or total category merger or stability of the existing cue structure, depending on exact nature of the transmission bias.

Conclusions

Sound change resulting from a cognitive restructuring of phonetic cue weights may be modeled as an adaptive strategy of probabilistic enhancement interacting with systematic biases in speech production. Computational simulations show that such a restructuring may come about without appealing to either (a) inherent perceptual bias for or against any particular cues or (b) a system-wide pressure or preference for contrast maintenance. Given just the initial state and characterization of transmission bias, this model allows us to make (probabilistic) predictions about directionality in sound change. Ongoing extensions of this work include experimental testing of the model predictions using human subjects.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Clayards, M. (2008). *The ideal listener: making optimal use of acoustic-phonetic cues for word recognition*. Unpublished doctoral dissertation, University of Rochester.

Clayards, M., Tanenhaus, M. K., Aslin, R., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*(1), 37-58.

Hyman, L. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic studies presented to Joseph H. Greenberg.* Saratoga: Anma Libri.

Kang, K.-H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of the Acoustical Society of America*, *124*(6), 3909–3917.

Lisker, L. (1978). Rapid vs. rabid: a catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research SR-54*, 128–32.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Toscano, J., & McMurray, B. (2008). Using the distributional statistics of speech sounds for weighting and integrating acoustic cues. In *Proceedings of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.