What do you expect from an unfamiliar talker?

One of the longest-standing puzzles in speech perception is how listeners cope with the often extreme differences in how individual talkers use acoustic cues to realize their linguistic intentions. A number of solutions have been proposed, including proposals that listeners quickly *adapt* to unfamiliar talkers by *learning* the distributions of acoustic cues that they produce (their "accent").

This can be formalized as a kind of statistical inference, where listeners try to infer which of all possible accents best explains a talker's speech (Kleinschmidt & Jaeger, 2015). In this view, prior experience with other talkers can help because it narrows down the range of possibilities that a listener needs to consider (in Bayesian jargon, it provides an *informative prior* on accents). We test a critical prediction of this view: when an unfamiliar talker's accent falls *outside* the range of typical variation across talkers, listeners should adapt only partially. Specifically, listeners' phonetic classifications should reflect a compromise between listeners' prior expectations and the actual accent they hear. We also, in doing so, demonstrate a novel technique for measuring listeners' subjective prior expectations about an unfamiliar talker's accent. Critically, this technique does not require the laborious collection and annotation of large quantities of speech from different talkers.

In a /b/-/p/ distributional learning paradigm (Clayards, Tanenhaus, Aslin, & Jacobs, 2008), listeners (n = 138; approx. uniformly distributed across conditions) hear a bimodal distribution over voice onset time (VOT), with a cluster at a low value implicitly corresponding to /b/ and another at a high value corresponding to /p/ (Figure 1). By varying the location of these clusters, we create accents that are more or less like those produced by a typical American English talker (as measured by, e.g., Kronrod, Coppess, & Feldman, 2012) (Figure 1).

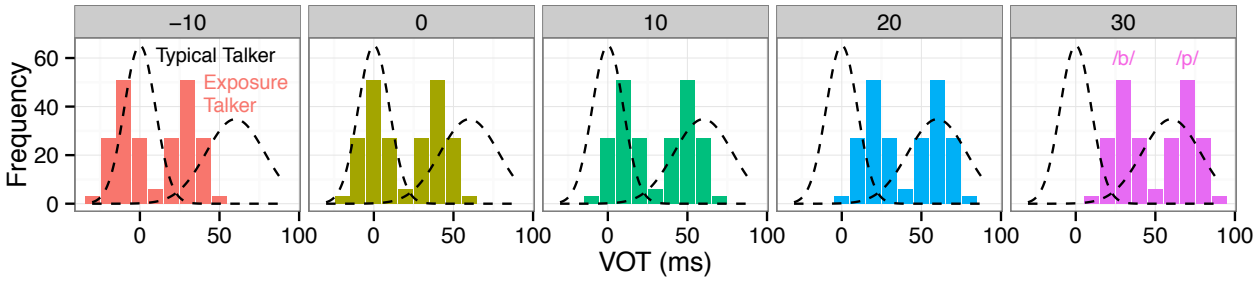

Figure 1: VOT distributions for each accent.

We measure how well listeners *learn* these accents by comparing their classification functions to the ideal boundaries implied by the exposure distributions alone (Figure 2). As predicted, when the VOT clusters were unusually high or low, listeners' *actual* category boundaries reflected a compromise between the typical (expected) talker and the statistics of the exposure talkers (see caption of Figure 2).

Second, we used a belief-updating model to work backwards from the patterns of adaptation to different accents, inferring what listeners' starting beliefs were (Figure 3), and how confident they were in those beliefs. The inferred prior expectations matched the range of typical American English talkers' /b/ and /p/ distributions, including the—at first blush—counterintuitive finding that listeners were *more* uncertain about the /b/ mean VOT than /p/. Even though the *within*-talker variance of VOT is higher for /p/ than /b/, the *between*-talker variance for /b/ is likely higher because some talkers pre-voice their /b/s with large, negative mean VOTs (Lisker & Abramson, 1964).

The ability to reverse-engineer listeners' prior expectations from perceptual data potentially provides an important and heretofore missing tool in the toolbox of laboratory phonology. Measuring these beliefs is important for at least two reasons. First, the task of learning a new talker's accent would be nearly as hard as learning the language for the first time were it not for listeners' prior experience with *other* talkers. Thus, understanding the remarkable ability of listeners to robustly comprehend speech from many different

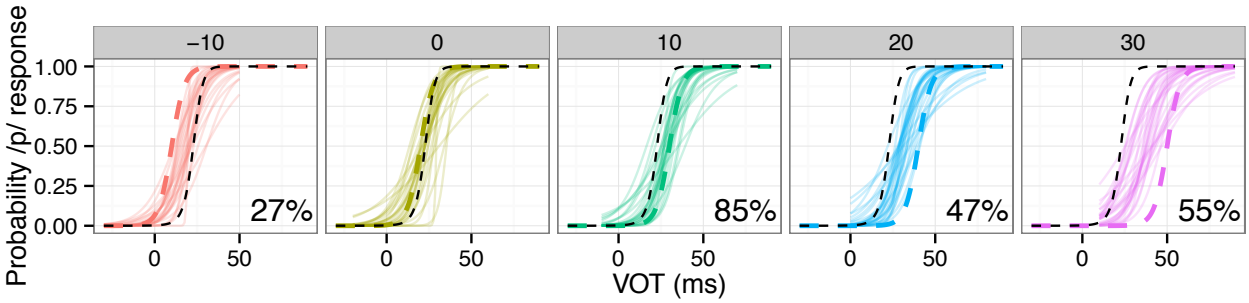

Figure 2: After exposure, listeners' /b/-/p/ classifications (thin colored lines; estimated via logistic GLM) reflected a compromise between the boundary for a typical talker (dashed black) and the experimental talker (dashed colored). Moreover, more extreme shifts on average led to less complete adaptation (lower percentage of boundary shift from typical to experimental; not shown for 0ms shift because experimental and typical talkers are too similar to reliably measure percentage).

talkers requires understanding the expectations they bring to an unfamiliar talker. Second, and relatedly, this technique can directly link the variability in *production* of linguistic variables with listeners' subjective expectations about those variables, both conditioned on *social* variables. Our proof-of-concept here (implicitly) uses standard American English, but the same procedure can be applied to specific variables like gender, region, class, etc., by providing information to the listener about *who* the talker is (which listeners do use to guide speech perception, Hay & Drager, 2010; Niedzielski, 1999; Strand & Johnson, 1996). We discuss the opportunities this offers for future research on speech perception and production.

References

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi:10.1016/j.cognition.2008.04.004

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. doi:10.1515/ling.2010.027

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi:10.1037/a0038695

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A Unified Model of Categorical Effects in Consonant and Vowel Perception. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 629–634). Austin, TX: Cognitive Science Society.

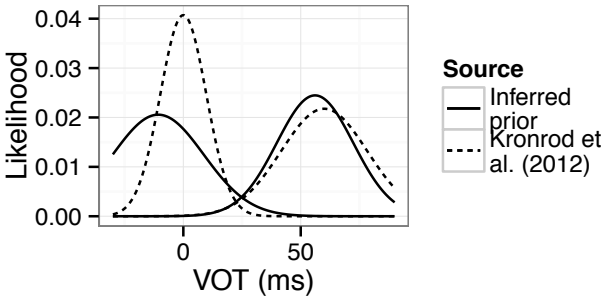

Figure 3: Cue distributions with maximum prior probability as inferred from adaptation data, compared with the distributions measured by Kronrod et al. (2012).

Lisker, L., & Abramson, A. (1964). A cross-language sudy of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. doi:10.1177/0261927X99018001005

Strand, E. A., & Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In D. Gibbons (Ed.), *Natural language processing and speech technology: Results of the 3rd kONVENS conference, bielfelt* (pp. 14–26). Berlin: Mouton de Gruyter.