What do you expect from an unfamiliar talker?

Dave F. Kleinschmidt[1], and **T. Florian Jaeger**[1,2,3]
{dkleinschmidt, fjeager} @ bcs.rochester.edu
[1]Department of Brain and Cognitive Sciences, [2]Department of Computer Science, and [3]Department of Linguistics,
University of Rochester, Rochester, NY, 14607 USA

Abstract

Keywords: Cognitive Science, Linguistics, Psychology, Language understanding, Learning, Speech recognition

Introduction

A longstanding problem in speech perception is how listeners manage to cope with substantial differences in how individual talkers produce speech. Recent evidence suggests that one strategy listeners employ is to *rapidly adapt* to unfamiliar talkers (Bertelson, Vroomen, & de Gelder, 2003; Clarke & Garrett, 2004; Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003, among others). Kleinschmidt and Jaeger (2015) showed that such adaptation can be modeled as a form of statistical inference. Each talker's particular accent (way of talking) can be formalized as the distribution of acoustic cues that they produce for each phonetic category (or other underlying linguistic unit). When framed this way, listeners can adapt to an unfamiliar talker by inferring what those cue distributions look like, based on the cues that are actually produced by the talker.

One insight that this *ideal adapter* framework provides is that rapid adaptation to an unfamiliar depends just as much on a listener's prior experience with other talkers as it does on the speech produced by the unfamiliar talker themself. A listener's experience with others talkers provides the starting point for the distributional learning required for adaptation, or, in Bayesian terms, a *prior distribution* over possible accents (cue distributions). More informative prior beliefs can substantially reduce the amount of direct evidence needed to converge on accurate beliefs about the current talker's cue distributions.

In this study, we test a critical predcition of this framework. To the extent that a listener's prior beliefs are informative, they must take some probability *away* from unlikely accents. Confronted by a talker whose accent falls well outside the range of what they expect based on their previous experience, the ideal adapter framework predicts that a listener will require more evidence to adapt, leading to slowed or incomplete adaptation. There is some limited evidence that this is the case. For instance, Idemaru and Holt (2011) found that listeners have difficulting adapting to a talker who produces anti-correlated distributions of two cues that are typically positively or un-correlated. Sumner (2011) found that listeners had trouble adapting to a talker who produced a distribution of cues for the /b/ and /p/ sounds that were substantially lower than a typical talker.

However, no studies have systematically probed whether and how a listener's prior expectations contrain rapid adaptation. To that end, we expose listeners to a range of different accents, which differ in the cue distributions for /b/ and /p/. By parametrically manipulating these distributions, we create a range of accents that are more or less similar to what a typical talker of English produces. We additionally sought to test a further predition. To the extent that a listener's prior expectations do, in fact, constrain their adaptation, it would be possible to *infer* what the underlying prior belief distributions are. This would provide a powerful method for investigating listeners' subjective prior beliefs that could be applied to other cues, categories, and even social variables (like gender, native language background, etc.).

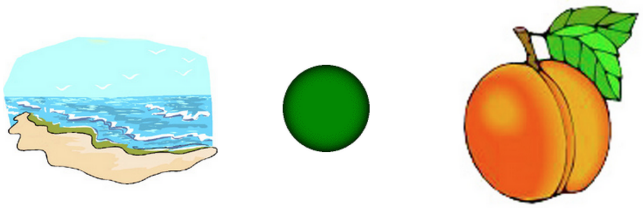

Figure 1: Example trial display (beach/peach). Listeners first click on the green button to play the word, then click on one picture to indicate what they heard.

Experiment

Methods

Subjects We recruited 169 subjects via Amazon's Mechanical Turk, who were paid $2.00 for participation, which took about 20 minutes. We excluded subjects who participated more than once ($n = 4$) or whose accuracy at 0 ms and 70 ms VOT—as extrapolated via a logistic GLM—was less than 80% correct ($n = 28$; $n = 1$ for both reasons). After these exclusions, data from 138 subjects remained for analysis for analysis.

Procedure Our distributional learning procedure is described in Kleinschmidt, Raizada, and Jaeger (2015), and is based on Clayards, Tanenhaus, Aslin, and Jacobs (2008). On each trial, two response option images appeard, which corresponded to one of three /b/-/p/ minimal pairs (beach/peach, bees/peas, or beak/peak). Subjects then clicked on a central button which played the corresponding minimal pair word, and then clicked on the picture to indicate whether they heard the /b/ or /p/ member of the minimal pair. Subjects performed 222 of these trials.

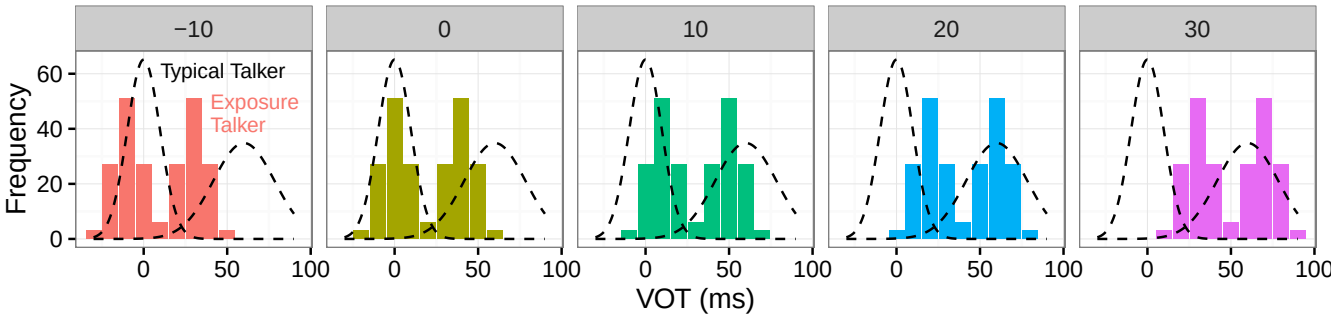

Figure 2: Each subject heard one of these five synthetic accents, which differ only in the distribution of VOTs of the word-initial stops. Black dashed lines show VOT distributions from a hypothetical typical talker (as estimated by Kronrod et al., 2012). Note that the 0 and 10ms shifted accents are reasonably close to this typical talker, while the −10, 20, and 30ms shifted accents deviate substantially.

Each trial's word was synthesized with a voice onset time (VOT) that was randomly drawn from a bimodal distribution, with low and high VOT clusters implicitly corresponding to /b/ and /p/, respectively. This distribution defined the *accent* that the subject heard, and each subject was pseudorandomly assigned to one of five accent conditions (Figure 2).

Results and Discussion

/b/ mean VOT	% boundary shift	95% CI (bootstrapped)
-10	29	13–44
0	-28	-107–51
10	80	57–104
20	53	45–62
30	49	40–57

Table 1: Percentage of boundary shift from typical talker to each exposure talker, averaged over subjects with 95% bootstrapped confidence intervals (compare with Figure 3). 0% shift corresponds to no adaptation at all, while 100% corresponds to perfect adaptation, ignoring any prior beliefs.

Figure 3 shows the classification functions for each individual listener. In each accent, these classification functions tend to fall in between the boundaries predicted by the typical talker distributions and the boundaries implied by the exposure distributions. We can quantify this by the percentage of the predicted shift in category boundary from the classification function for the typical talker to the boundary implied by the input distribution (Table 1). A 0% shift corresponds to no adaptation at all, while a shift of 100% corresponds to complete adaptation to the exposure distributions, with no (remaining) influence of the typical talker.

In all conditions, the average shift percentage was between 0% and 100% (except the 0ms shift condition, which is so close to the typical talker that estimating the percentage is numerically unstable). More interestingly, the more extreme conditions show less complete adaptation than the less extreme conditions. Together, these results suggest that listeners' adaptation was constrained by their prior expectations (given the finite amount of evidence they received about the unfamiliar talker). This provides qualitative evidence that listeners combine their prior expectations with observed cue distributions in order to rapidly adapt to unfamiliar talkers, as predicted by the ideal adapter framework (Kleinschmidt & Jaeger, 2015).

Model

Our second goal in this paper is to test whether it is possible to infer listeners prior beliefs based on their patterns of adaptation to different accents. To that end, we use a variant of a Bayesian belief-updating model that has previously provided a good account of how listeners incrementally update their beliefs in order to rapidly adapt to an unfamiliar talker (Kleinschmidt & Jaeger, 2011, 2012, 2015). Previous modeling work along these lines has treated the content of listeners prior beliefs—the category means and variances they think are most likely—as known and fixed, setting them based on pre-adaptation classification data, and then fitting the confidence in those prior beliefs as a free parameter. Here, we wish to fit both the content of (prior expected mean and variance of each category) and the confidence in prior beliefs, based on adaptation data presented above.

Methods

We denote the listener's beliefs about the *current* talker's generative model as the parameters of a two-component mixture of gaussians

$$\theta = \{\mu_b, \sigma_b^2, \mu_p, \sigma_p^2\} \quad (1)$$

where μ is a category's mean VOT and σ^2 is its variance. As in Kleinschmidt and Jaeger (2015), we use an independent, conjugate Normal-χ^2 prior for each category, with parameters (Gelman, Carlin, Stern, & Rubin, 2003)

$$\phi = \{\mu_{0,b}, \sigma_b^2, \mu_{0,p}, \sigma_p^2, \kappa_0, \nu_0\} \quad (2)$$

$$\theta|\phi \sim \prod_{c\in\{b,p\}} \text{Normal}(\mu_c|\mu_{0,c}, \frac{\sigma_c^2}{\kappa_0})\chi^{-2}(\sigma_c^2|\sigma_{0,c}^2, \nu_0) \quad (3)$$

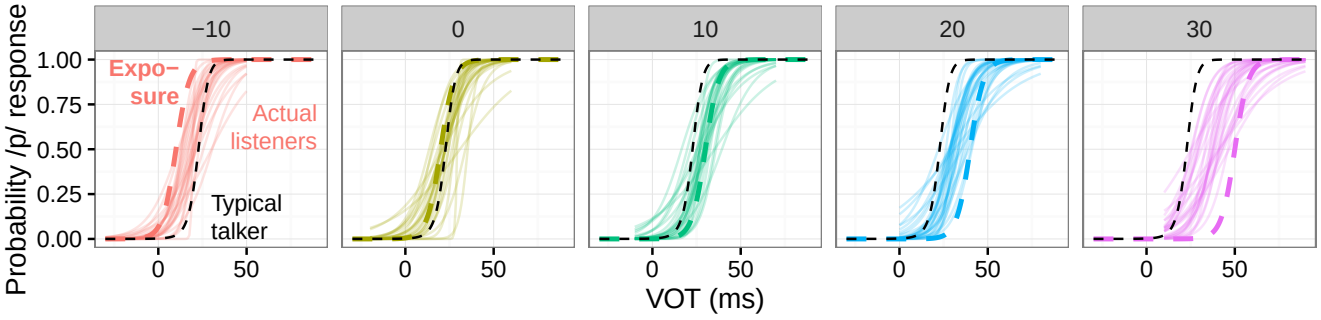

Figure 3: Listeners' responses, smoothed with logistic functions (thin lines), compared with the classification functions expected based on a typical talker (no learning; dashed black lines) and complete adaptation to the exposure distributions (thick dashed colored lines). Listeners' actual category boundaries lie between the typical talker and exposure talker boundaries (see Table 1).

where μ_0 and σ_0^2 are a category's prior expected mean VOT and variance, respectively, and κ_0 and ν_0 are the listener's confidence in these prior expectations, measured as pseudocounts. Note that, as in previous modeling work in this framework, the prior confidence parameters are shared between the two categories. Preliminary simulations showed that it wasn't possible to uniquely identify the model using separate prior confidence parameters for the two categories.

To estimate the listeners' prior beliefs, we infer values for these parameters given the observed adaptation behavior (category responses y and input VOTs x) using Bayesian inference, marginalizing over θ:

$$p(\phi|x,y) \propto p(y|\phi,x)p(\phi) \quad (4)$$

$$\propto \int d\theta p(y|\theta,x)p(\theta|x,\phi)p(\phi) \quad (5)$$

We make the simplifying assumptions that the order of the trials does not matter (exchangeability), and that for the purposes of belief updating ($p(\theta|x,\phi)$) the category labels that we, the experimenters, assigned to each cluster are known. This is equivalent to assuming that listeners pick up on the cluster structure of the input they receive, accurately detecting the mean and variance of each cluster, and finding a compromise between these observed statistics and their prior beliefs, possibly revising earlier decisions they made about assigning stimuli to one category or the other. Backing off this assumption is possible but computationally trickier, and we leave it as a question for future work. It does not, as far as we can determine, bias our results one way or the other. Finally, we also add a small lapsing rate parameter, that allows for some proportion of responses to be attributed to random guessing (see Clayards et al., 2008 for a discussion).

The posterior distributions of each of these parameters (the shared prior beliefs plus lapsing rate) were estimated using MCMC with the Stan software package (Stan Development Team, 2015). Weakly informative hyperpriors were used that were centered at 0 with standard deviations of 100 for the prior expected means and variances (making them roughly constant over reasonable values) and 888 (four times the total number of trials that listeners heard) for the prior confidence pseudocounts (covering the whole range from completely ignoring the prior to never adapting at all). The prior for the lapsing rate was uniform on $[0,1]$. We ran four chains for 1000 samples each, discarding the first 500 as burn-in for a total of 2000 samples overall. This sampler converged well and achieved good mixing (maximum Gelman-Rubin $\hat{R} = 1.01$; Gelman & Rubin, 1992).

Results

The first way we evaluate this model is to ask how well it fits listeners' behavior. Figure 4 shows listeners' average classification functions, compared with the posterior predictive classification functions from the belief-updating model. The first thing to notice is that the model fits the data well (log-likelihood ratio vs. an intercept-only binomial null model of 12385 and Spearman's $\rho = 0.9$) capturing the different classification functions that result from exposure to each input distribution. This in and of itself is an interesting result: it shows that there does exist some set of prior beliefs such that the range of adaptation behavior we observed can be explained by a model where all listeners start from a common set of prior beliefs.

Parameter	Expected	95% HPD Int.	Units
κ_0	243	163–495	observations
ν_0	772	493–1180	observations
$\mu_{0,b}$	-11	-28–3	ms VOT
$\mu_{0,p}$	56	44–72	ms VOT
$\sigma_{0,b}$	19	16–23	ms VOT
$\sigma_{0,p}$	16	14–20	ms VOT

Table 2: Expected values and 95% highest posterior density intervals for the prior parameters, given the adaptation data.

The second way to evaluate this model is based on the prior beliefs it infers listeners to have. Table 2 shows the posterior

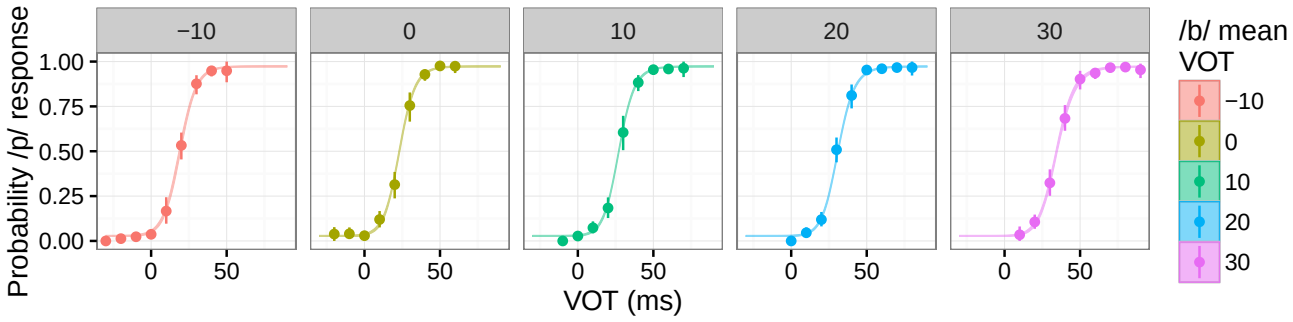

Figure 4: The classification functions (shaded ribbons, 95% posterior predictive intervals) predicted by the belief updating model fit listeners' responses well (dots with lines showing bootstrapped 95% confidence intervals).

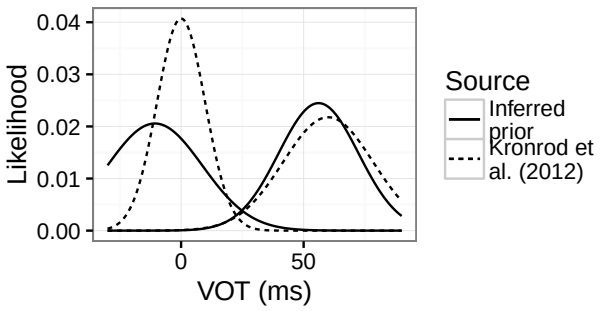

Figure 5: Expected cue distributions based on inferred prior beliefs from adaptation data (solid lines). Plotted with VOT distributions measured by Kronrod et al. (2012) based on a combination of classification and discrimination behavior (dashed lines).

expectation and 95% highest posterior density intervals for each of the prior belief parameters given the adaptation data above. The behavioral data was consistent with high confidence in prior beliefs with the prior confidence about category variances ($E(\nu_0) = 772.39$) higher than confidence in the category means ($E(\kappa_0) = 243.26$). Both of these (measured in pseudo-observations) are larger than the number of trials that listeners heard in the experiment (222), which means that as far as the belief-updating model is concerned, listeners updated beliefs reflected their prior belies as much as (in the case of the means) or more than (for the variances) the distributions they actually observed. This is consistent with the qualitative finding that listeners' category boundaries are intermediate between the boundaires corresponding to a typical talker and the experimental exposure talker.

Figure 5 shows the cue distributions corresponding to the posterior expected values of the prior expected mean and variance parameters given the behavioral data, compared with the distributions corresponding to a typical talker (as determined by a combination of classification and discrimination data, Kronrod et al., 2012, see also Lisker & Abramson, 1964). The inferred prior beliefs are in reasonably good agreement with this typical talker, with one excpetion: the mean for /b/ is slightly lower ($E(\mu_{0,b}) = -10.69$ ms), and the standard deviation of /b/ is slightly higher ($E(\sigma_{0,b}) = 19.4$ ms).

Discussion

These modeling results show that the pattern of adaptation behavior observed above is consistent with a belief-updating model of phonetic adaptation that combines prior expectations with input statistics in order to infer the current talker's cue distributions. Specifically, it shows that there exists a single set of prior beliefs that captures the range of adaptation to different input distributions that stretches from nearly complete adaptation to partial adaptation at best.

These prior beliefs are reasonably consistent with other attempts to determine what listeners think the underlying cue distributions are (Kronrod et al., 2012), as well as the distributions produced by actual talkers (Lisker & Abramson, 1964). In fact, the prior expected VOT distribution for /p/ that our model inferred is almost identical to that observed by both Kronrod et al. (2012) and Lisker and Abramson (1964). The distribution for /b/ deviates from prior work, however. One possible reason for this is that a substantial minority of English speakers produce prevoiced /b/ (Lisker & Abramson, 1964; Docherty, Watt, Llamas, Hall, & Nycz, 2011), which is characterized by a lower (negative) VOT and a higher variance (often higher even than /p/). That is, across talkers, the /b/ VOT distribution parameters (mean and variance) have a *bimodal* distribution. We assumed a single, unimodal prior distrbution, and the prior beliefs we inferred to be most likely are consistent with a compromise between the two types of /b/ distributions that talkers actually produce.

This possibility suggests two directions for future work. First, large-scale corpus studies of VOT distributions would be informative about the true underlying cross-talker variability in these distributions, but the only study of this type we are aware of considers only non-negative VOTs (Chodroff, Godfrey, Khudanpur, & Wilson, 2015) and thus excludes talkers who prevoice their /b/. Second, more modeling work is needed to test whether a multimodal prior is justified given the adaptation data, and if so, whether it would change the inference about listeners' prior expectations for /b/.

The other major result of this modeling is that listeners

have high confidence in their prior expectations about the VOT distributions of /b/ and /p/, acting as if they had already observed around 200–800 samples from each category (for the category means and variances, respectively) from the unfamiliar talker they encountered in our experiemnt. These confidence values are one or two orders of magnitude higher than those inferred in other belief-updating modeling (Kleinschmidt & Jaeger, 2015).

However, this previous work was based on adaptation to a /b/-/d/ contrast, which is cued by spectral cues (formant frequency transitions) which generally vary substantially across talkers (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson, 1952) . The acoustic cues to the /b/-/p/ contrast used in the current study do not show as much variability across talkers (Allen, Miller, & DeSteno, 2003; Chodroff et al., 2015). When there is little variability across talkers, past experience with other talkers' VOT distributions is highly informative about the distributions that an unfamiliar talker will produce, requiring less adaptation. Likewise, when there is more variability across talkers, listeners need to rely more on the current talker's cue distributions and less on their prior experience. Thus, the apparent discrepancy between the confidence that listeners place in their prior beliefs in the current study and in Kleinschmidt and Jaeger (2015) is entirely consistent with an "ideal adapter" which combines prior beliefs with current experience weighted according to confidence (Kleinschmidt & Jaeger, 2015). This idea finds further empirical support in Kraljic and Samuel (2007), who found that after the same amount of exposure, listeners recalibrate a /d/-/t/ contrast (analogous to the /b/-/p/ contrast used here) much less than an /s/-/ʃ/ contrast (where there is substantial variability across talkers; McMurray & Jongman, 2011; Newman, Clouse, & Burnham, 2001).

Conclusion

A central prediction of the ideal adapter framework (Kleinschmidt & Jaeger, 2015) is that listeners adapt to unfamiliar talkers by combining their prior beliefs with observed evidence about that particular talker's cue distributions. In this paper, we have shown first that for a range of different accents (cue distributions), listeners' behavior in a distributional learning experiment reflects a compromise between what would be expected for the cue distributions produced by a typical talker and the exposure talker. Second we have shown that the range of adaptation behavior observed across the various accents listeners heard can be captured by a single belief-updating model, which assumes that listeners start from shared prior expectations and update them based on experience with the exposure talker.

These results emphasize the importance of listeners' prior expectations for robust speech perception in the face of talker variability. Even if all the listener knows about the talker is that they are speaking English, they can still benefit from prior experience with other speakers of English to provide an informative head start for adaptation. The modeling framework we use has the additional advantage of allowing us to *infer* what cue distributions listeners believe an unfamiliar talker will produce. This provides a potentially powerful—and heretofore missing—tool for probing listeners' prior expectations, based only on comprehension data. These beliefs reflect what listeners have learned about the variability they can expect across talkers, and probing how this internal model is related to the *actual* variability across talkers (measured via speech production data) is an important next step in advancing our understanding of robust speech perception.

More generally, prior knowledge is increasingly understood to play in important role in a number of perceptual and memory domains (e.g., Brady, Konkle, & Alvarez, 2009; Brady & Tenenbaum, 2013; Froyen et al., 2015; Orhan & Jacobs, 2011). Distributional learning provides an approach to probing prior expectations about the *statistics* of the sensory world, which, as in speech perception, are critical to effectively coping with non-stationarity in sensory statistics.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544. doi: 10.1121/1.1528172

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Brady, T. F., Konkle, T., & Alvarez, G. a. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502. doi: 10.1037/a0016797

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85–109. doi: 10.1037/a0030779

Chodroff, E., Godfrey, J., Khudanpur, S., & Wilson, C. (2015). Structured Variability in Acoustic Realization : A Corpus Study of Voice Onset Time in American English Stops. In T. S. C. f. I. 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences.* Glasgow, UK: the University of Glasgow.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647. doi: 10.1121/1.1815131

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

Docherty, G., Watt, D., Llamas, C., Hall, D., & Nycz, J. (2011). Variation in Voice Onset Time Along the Scottish-English Border. In W.-S. Lee & E. Zee (Eds.), *Proceedings*

of icphs xvii (pp. 591–594). International Phonetic Association.

Froyen, V., Feldman, J., Singh, M., Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian Hierarchical Grouping : Perceptual Grouping as Mixture Estimation Bayesian Hierarchical Grouping : Perceptual Grouping as Mixture Estimation. *Psychological Review*, *122*(4), 575–597.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second ed.). Taylor & Francis.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5.1), 3099–111.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–56. doi: 10.1037/a0025641

Kleinschmidt, D. F., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd acl workshop on cognitive modeling and computational linguistics.* Stroudsburg, PA: Association for Computational Linguistics. Talk.

Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 605–10). Austin, TX: Cognitive Science Society. Talk.

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

Kleinschmidt, D. F., Raizada, R., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. In R. Dale et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. doi: 10.1016/j.jml.2006.07.010

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.

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

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–46. doi: 10.1037/a0022325

Newman, R. S., Clouse, S. a., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196. doi: 10.1121/1.1348009

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi: 10.1016/S0010-0285(03)00006-9

Orhan, A. E., & Jacobs, R. A. (2011). A Nonparametric Bayesian Model of Visual Short-Term Memory. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2451–2456). Austin, TX: Cognitive Science Society.

Peterson, G. E. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175. doi: 10.1121/1.1906875

Stan Development Team. (2015). *Stan: A C++ Library for Probability and Sampling, Version 2.9.0.*

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–6. doi: 10.1016/j.cognition.2010.10.018