Learning to understand an unfamiliar talker:
Testing models of adaptive speech perception

Maryann Tan[1, 2] & T Florian Jaeger[2, 3]

[1]Centre for Research in Bilingualism, University of Stockholm; [2]Department of Brain & Cognitive Sciences, University of Rochester; [3]Department of Computer Science, University of Rochester

Index Terms: speech perception, distributional learning, adaptation, error-driven learning

1. Introduction

Human speech perception is a computational feat. Critical to this ability is *adaptivity*: a few minutes of exposure can significantly reduce the processing difficulty associated with initial encounters with unfamiliar accents. How such adaptation unfolds incrementally, however, remains largely unknown, leaving basic predictions by theories of adaptive speech perception untested. This includes questions about how listeners' prior expectations based on lifelong experiences are integrated with the unfamiliar speech input, as well as questions about the speed and success of adaptation.

We begin to address these knowledge gaps in a novel incremental exposure-test paradigm aimed at testing core predictions of distributional learning models of speech perception [1-3]. Specifically, that:

1) The direction and magnitude of change depend monotonically on listeners' prior experience relevant to the present task.
2) The direction and magnitude of change depend on:
    a) the amount of exposure to the unfamiliar talker
    b) the distribution of phonetic cues of the unfamiliar talker.
3) Adaptation proceeds until full convergence on the statistics of the unfamiliar talker.
4) Adaptation proceeds quickly and slows down prior to convergence.

2. Methods

We expose 3 groups (N=119) of L1-US English listeners to different degrees of shifted phonetic distributions of VOT, the main cue for distinguishing word-initial stops (e.g., "dill" vs. "till"), while incrementally assessing cumulative changes in listeners' perception. The 3 conditions were identical in distribution except for the location of the predicted PSEs (point of subjective equality)—i.e., the point at which listeners are equally likely to respond "d" or "t"— set at 25, 35 or 65 msec. We use Bayesian mixed-effects psychometric models to characterize listeners' behavioral changes. Focusing on changes in listeners' PSEs, we compare changes in PSE against those predicted by idealized learners (ideal observers that perfectly learn the statistics of the input) and a model of adaptive speech perception (ideal adaptors that learn and infer those statistics through integration with their prior beliefs).

3. Results

Our findings confirm prediction 1, 2a and 2b; demonstrating gradient incremental adaptation that takes into account prior experience (Fig. 1). Comparisons of listener behavior across test blocks however show that the rate of incremental change decreased with increasing exposure (prediction 4), suggesting previously undocumented (but plausible) limits to adaptation. Our results pose a challenge to existing distributional learning models of speech perception which as yet do not predict constraints on adaptation.

We discuss potential extensions of exemplar and Bayesian inference models that can account for these findings. One of which refers to a distinction between *model learning* and *model selection* [4]: if at least the early moments of adaptive speech perception—those studied in our experiment—are limited to *selection* of previously experienced input, rather than learning of new distributional models, this could explain our results. We present initial—post-hoc—analyses that assess this possibility by comparing the limits in listeners' adaptivity to the range of talker-specific category boundaries that a typical L1-US English listener would have been likely to experience throughout their life (using a phonetic database of L1-US English [5]).

4. References

[1] Johnson, K., "The auditory/perceptual basis for speech segmentation.", WPL-OSU., vol. 50, pp. 101-113, Jul, 1997.

[2] Apfelbaum, K. S., McMurray, B., "Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization.", PBR, vol. 22, pp. 916-943, 2015.

[3] Sohuglu, E., Davis, M., "Perceptual learning of degraded speech by minimizing prediction error.", PNAS, vol.113, no.12, pp. E1747-E1756, Jan, 2016.

[4] Kleinschmidt, D., Jaeger, T. F., "Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel.", Psych. Rev., vol. 122, no.2, pp. 148-203, 2015.

[5] Chodroff, E., Wilson, C., "Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English.", Journ. Of Phon., vol. 61, pp. 30-47, 2017.

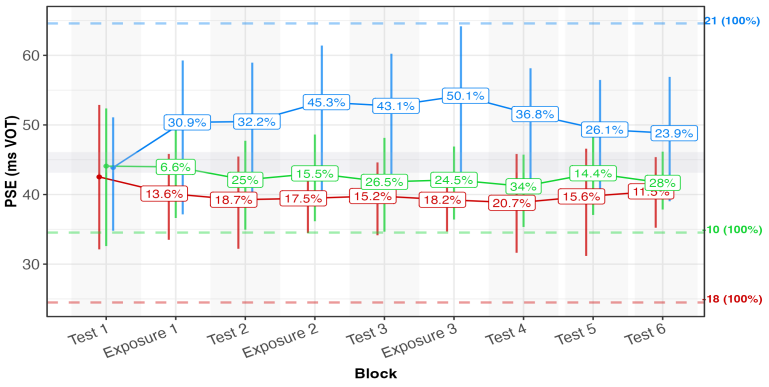

Figure 1. Summary of results. Changes across blocks and conditions in listeners' point-of-subjective-equality (PSE) of the lapse-corrected categorization functions. Point ranges represent the posterior medians and their 95%-CIs derived from the psychometric model. Horizontal dashed lines indicate 95%-CIs of the PSEs expected from an idealized learner (an ideal observer model that has fully learned the exposure distributions). Percentage labels indicate the degree of shift in the PSE exhibited by participants as a proportion of the expected shift under the idealized learners.