Speech perception, the lack of invariance, and adaptation: A computational level analysis

Dave F. Kleinschmidt & T. Florian Jaeger
University of Rochester, Department of Brain and Cognitive Sciences
Reprints:
dkleinschmidt@bcs.rochester.edu

Problem of lack of invariance: interpretation of acoustic cues varies across environments.

Proposed solution: listeners use a **generative model** to **predict** language input. **Prediction error** leads to adaptation (updating beliefs about the generative model)

Applies to predicting **environments**, too (what kind of talkers are expected)

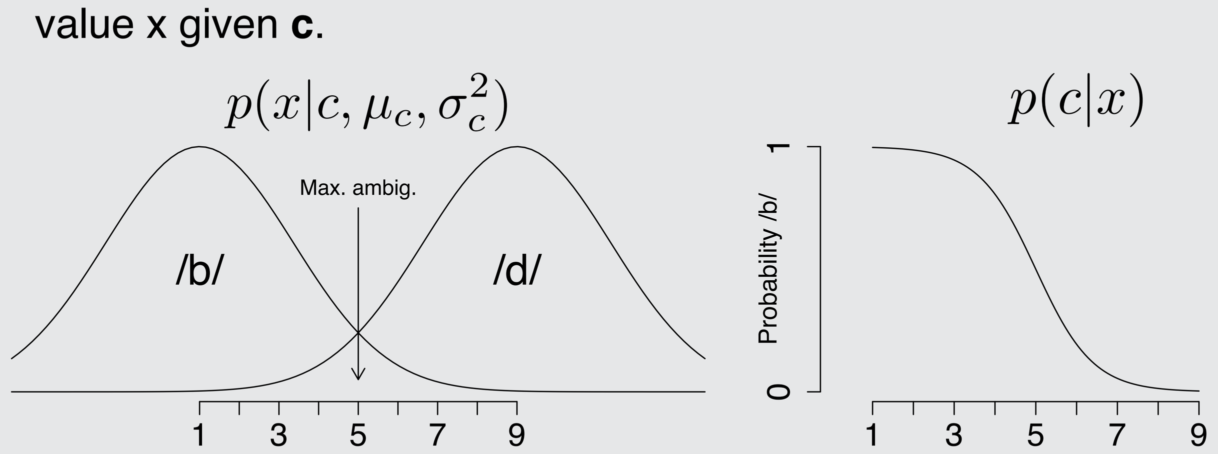
Provides a novel, unified perspective on adaptation in new environments, and generalization of adaptation across environments (talkers)

SPEECH PERCEPTION

- Goal: infer intent behind observable cues, via intermediate linguistic units (phonetic categories, words, syntactic structures, etc.)
- Uncertainty is present at every stage (ambiguity and noise)
- Optimal inference under uncertainty is described by Bayes Rule:

$$p(c|x) \propto p(x|c)p(c)$$

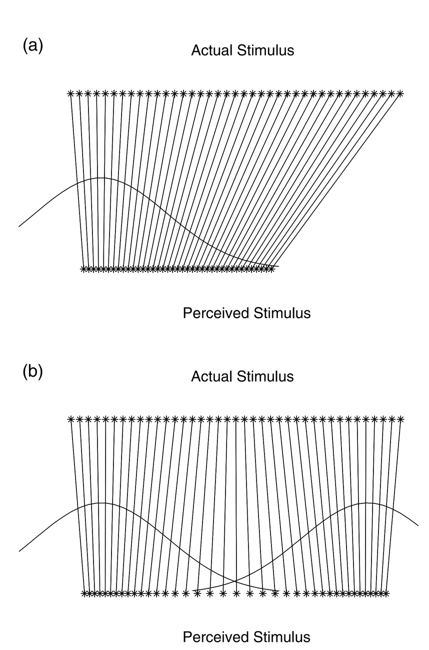
- Combines prior probability of c and likelihood of observing cue value x given **c**.



PERCEPTUAL MAGNET EFFECT
(Modeling: Feldman, Griffiths, and Morgan, 2009)

- Influence of categories pulls percept towards category mean
- Separate variability due to category variance and production/perception noise.
- Infer speaker's intended target cue value based on observed cue value and knowledge of distributions (category variance and noise)

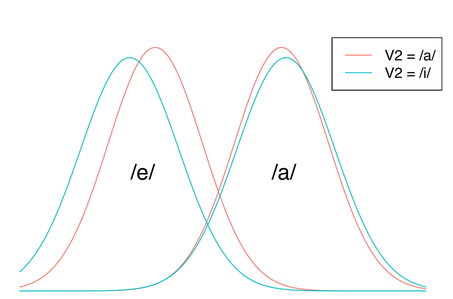
$$p(x_T|x_S) \propto \sum_c p(x_S|x_T, c)p(x_T|c)p(c)$$



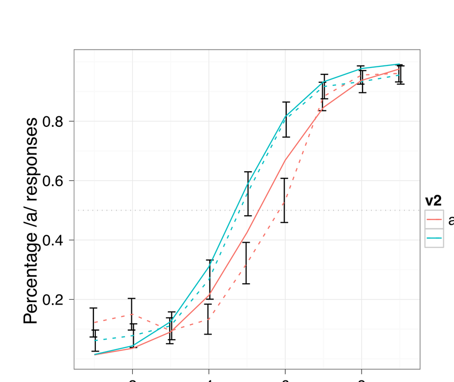
COMPENSATION FOR COARTICULATION
(Modeling: Sonderegger and Yu, 2010)

- Vowel-to-vowel coarticulation: first vowel takes on characteristics of following vowel.
- Listeners compensate for this by shifting their category boundaries.
- Model by conditioning likelihood for first vowel on second vowel

$$p(V_1|x, V_2) \propto p(x|V_1, V_2)p(V_1)$$



- Measure categorization responses to first vowel in bV1bV2 words (V1 is /a/ or /e/, V2 is /a/ or /i/).
- Compute cue distribution for each V1, V2 combination based on production data.



ADAPTATION

- Good comprehension depends on accurate likelihood $p(x|c, \mu_c, \sigma_c^2)$ (the distribution of cues for each category, characterized by mean and variance)
- Lack of invariance: likelihood changes across contexts due to differences in environments (speaker, dialect, etc.)
- A rational comprehension system is sensitive to these differences in distributions.

INCREMENTAL BELIEF UPDATING: Adapting to changes in the underlying distributions

- Don't have access to the "true" likelihood distribution, but uncertain beliefs about category parameters

$$p(\theta_c) = p(\mu_c, \sigma_c^2)$$

- Have to infer distributions (means and variances) and intended categories together:

$$p(\mu_c, \sigma_c^2, c|x) \propto p(x|\mu_c, \sigma_c^2, c)p(\mu_c, \sigma_c^2)p(c)$$

- Combine prior beliefs and current experience to do **incremental belief updating.**
- Compare predictions from beliefs with currently processed speech. Use prediction error to update.

RECALIBRATION (AND SELECTIVE ADAPTATION)

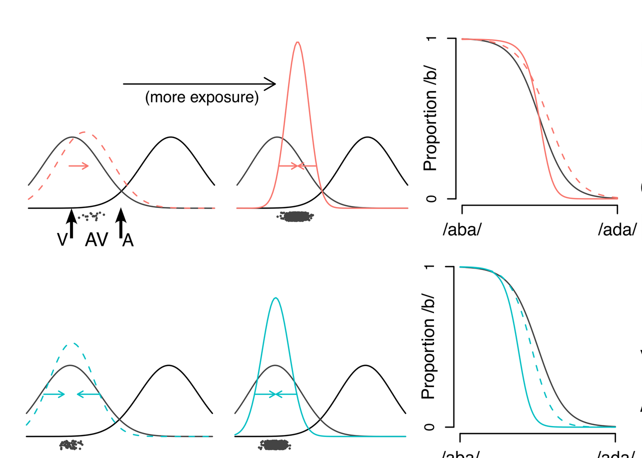
Behavior: Vroomen et al. (2007)

Recalibration: ambiguous acoustic cue (e.g. /b/-/d/) paired with disambiguating information (video of speaker producing /b/). More /b/ responses to audio-only test items, but effect fades with more cumulative exposure.

Selective adaptation: prototypical /b/ repeated many times. Fewer /b/ responses.

Modeling: Kleinschmidt & Jaeger (2011, 2012)

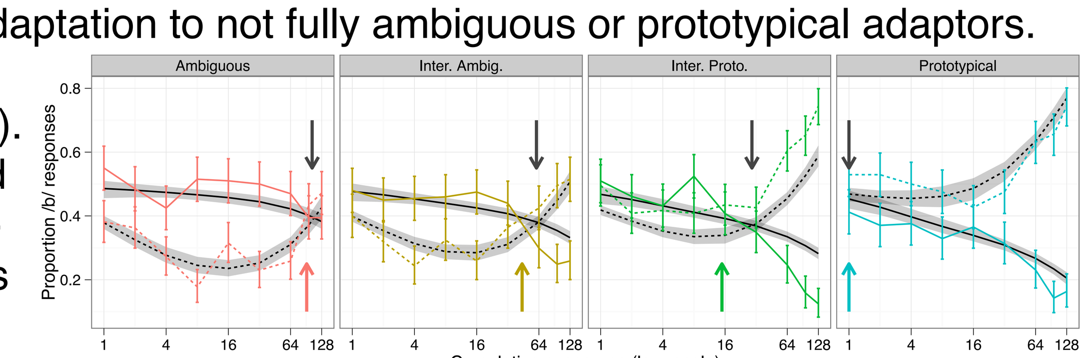
- Trial-by-trial adaptation predictions based on stimulus distribution:



McGurk effect creates not-fully-ambiguous /b/ percept. Initial mean shift results in more /b/ likelihood for acoustically intermediate test stimuli, but as variance decreases likelihood decreases as well for test stimuli, leading to decrease in recalibration effect.

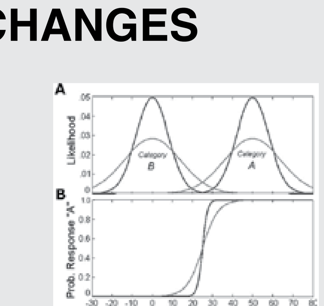
Repeated exposure to same prototypical /b/ results in lower variance and thus less /b/ likelihood for test stimuli and fewer /b/ responses

Predictions: intermediate adaptation to not fully ambiguous or prototypical adaptors. Replicated on MTurk (plus two intermediate conditions). Fit model to ambiguous and prototypical conditions; predict intermediate (fits just as well)



ADAPTED CATEGORY BOUNDARY DUE TO VARIANCE CHANGES
(Behavior+modeling: Clayards et al. 2008)

- Category boundary slope reflects uncertainty in classification
- Steeper for lower variance distributions
- Exposed listeners to low and high variance VOT distributions
- Found steep/shallow slopes, respectively.



RECAP

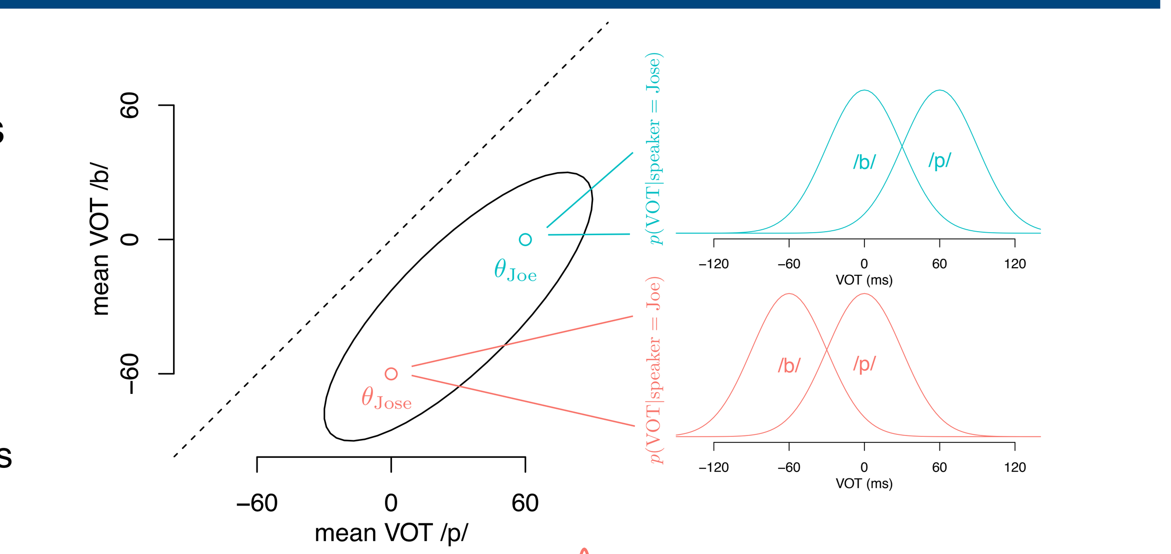
Previous work: **speech perception** (others) and **adaptation in novel environment** (us) as prediction/inference in a generative model.

Proposal: speech perception/adaptation **across speech environments** as prediction/inference in a generative model of **clusters of environments.**

GENERALIZATION

LOOKING FORWARD

- Speakers are characterized by the parameters of their category likelihoods $p(x|\mu_c, \sigma_c^2)$
- Prior beliefs about category parameters $p(\mu_c, \sigma_c^2)$ are really a prior over speakers



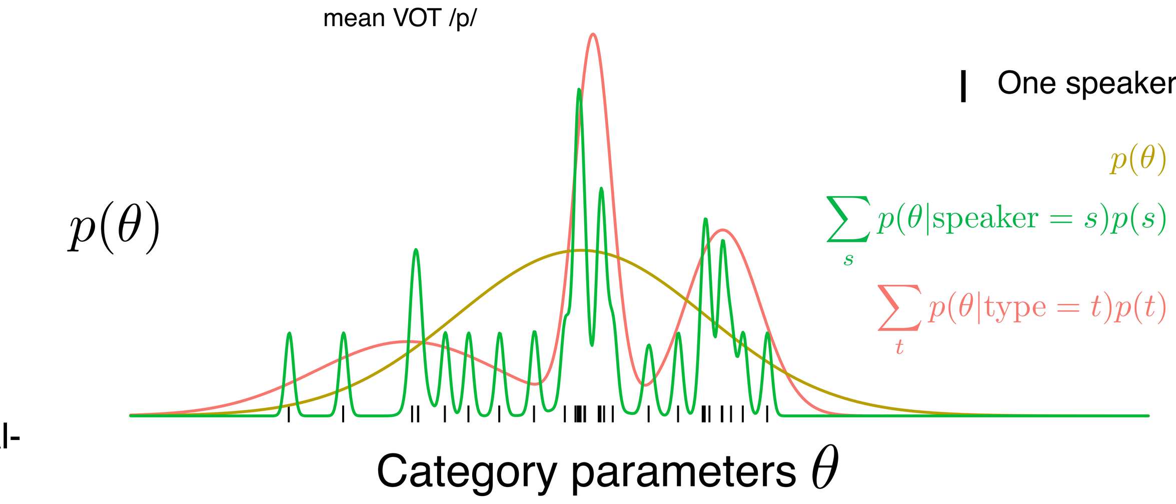
What priors should a rational learner have?

Prior should be *representationally efficient*, and depends on what kind of variability there is across environments (Anderson 1991):

Random variation: flat. Prior is weak, must re-adapt every time environment changes.

Variation due to different speakers: spiky. Prior is strong near familiar speakers (allows "swapping in" of right likelihood), and weak everywhere else.

Structured variation due to speaker groups: lumpy. Prior is strong, near highly familiar individuals (e.g. mom), and broader and less strong around similar-sounding groups (e.g. people with German accents). Allows *flexible generalization*.



There **is** structured variation among talkers (gender, accent, etc.)

The **optimal** prior is thus a **hierarchical:** clustered environments/talkers

This predicts:

- Rapid adaptation in new environments which are dissimilar from previously encountered ones (e.g. Norris et al. 2003; Vroomen et a. 2007; Kleinschmidt & Jaeger 2011, 2012)
- Robust adaptation that lasts (e.g. Eisner et al. 2006; Kraljic & Samuel 2005)
- Generalization depends on **similarity** with previous environments **and** expectation of new envi.

GENERALIZATION ACROSS SPEAKERS DEPENDS ON PRIOR EXPERIENCE:

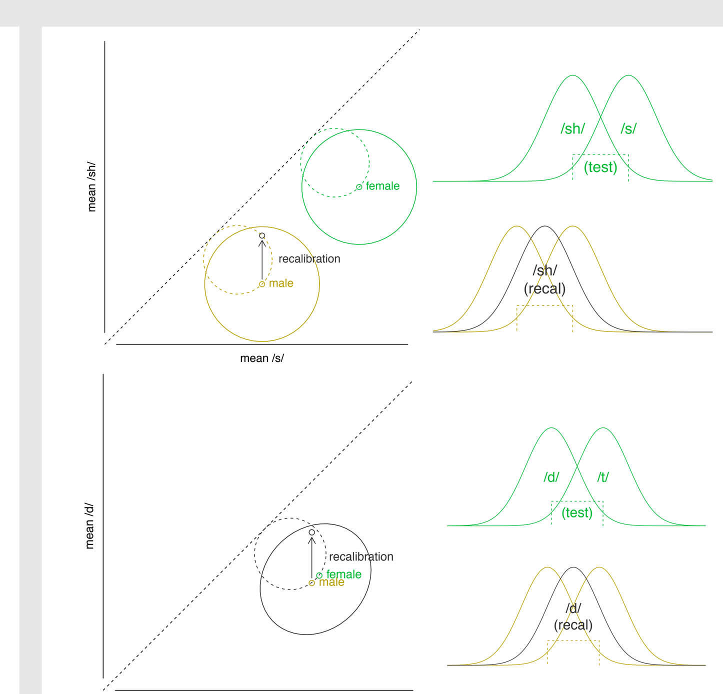
Generalization occurs when speakers are clustered together (use same set of updated beliefs). Listeners must infer clustering and speaker parameters on the fly.

GENERALIZATION IN RECALIBRATION
(Behavior: Kraljic & Samuel, 2007)

Recalibration of voicing (/d/-/t/) or fricative place (/s/-/sh/) contrast. Voicing generalizes from male to female talker but fricative does not.

Why?

- Male and female talkers differ systematically in fricative cues (spectral center of gravity), but not as much voicing cues (VOT).
- Listeners thus have strong **prior** that male and female speakers should not cluster together.
- Additionally, test stimuli have different acoustic cue ranges (low **likelihood** of shared cluster).



TALKER-INDEPENDENT ACCENT ADAPTATION
(Behavior: Bradlow & Bent, 2008)

Test comprehension on Mandarin-accented test talker after training with: 1) Same talker. Train on test talker. 2) Single talker. Train on different Mandarin-accented talker, 3) Multiple talker. Train on four different Mandarin-accented talkers (one quarter as much on each) Results: Same and multiple talker training both produce large gains in accuracy. Single talker is no better than task control.

Why?

- Single talker prior is peaked (high confidence) but wrong for the test talker. Either uninformative or misinformative.
- Multiple talker prior is broader but averages out idiosyncrasies of individual training talkers (and hence filters out misleading variation in test talker's speech).