1 Introduction

- What is the goal of speech perception?
- Infer speaker's intentions based on *noisy* and *ambiguous* signal.
- Accomplished by inferring intermediate linguistic representations: phonemes, words, phrases, syntactic structures, etc.
- Ambiguity and noisiness is present at every stage, so speech perception is still a problem of inference under uncertainty.
- Bayesian inference describes the optimal way to make use of partially informative cues in order to infer underlying linguistic categories that generated those cues.
- This inference depends on having a good *generative model* (knowing the distribution of cues associated with each category)
- This, we think, is the key insight to understanding adaptation.

1.1 A computational analysis of speech perception and adaptation

- Our approach is to develop a *computational level* analysis of speech perception (in the sense of Marr 1982 and Anderson, 1990).
- This analysis deals with the structure of the information which is available and the goals of the whole process
- This leads to a sense of what the optimal way of solving the task is, given the constraints imposed by the available information and the structure of the task.
- But doesn't (directly) address the *algorithms* by which speech perception is carried out, or how those *algorithms* are implemented.

2 Speech perception is probabilistic

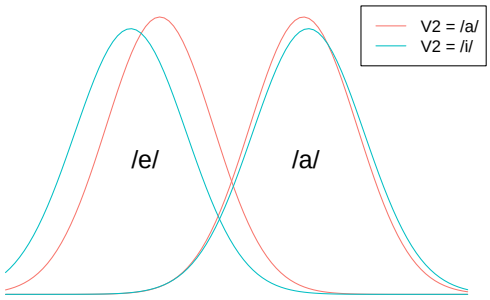
- Think of speech perception as inference of underlying categories, based on some cues.
- Want to compute the *posterior probability* that the category is b:

$$p(c = b|x) = \frac{p(x|c = b)p(c = b)}{\sum_c p(x|c)p(c)} \tag{1}$$

- Combines *likelihood*, $p(x|c = b)$, probability that cue value x observed given that intended category really is /b/, and *prior*, $p(c = b)$, the probability that /b/ occurs overall, regardless of the observation.

2.1 Perceptual magnet effect (Feldman, Griffiths, and Morgan, 2009)

2.2 Compensation for coarticulation (Sonderegger and Yu, 2010)



3 Lack of invariance, adaptation, and belief updating

- Problem of lack of invariance: the cues that are used to realize a category change across environments.
- Largely due to systematic variation, e.g. between different speakers or accents.
- Successful inference depends on having an accurate likelihood function $p(x|c)$, but this is exactly what changes across environments.
- For instance, male speakers tend to produce lower frequency formants than female speakers, so the true likelihood function $p(\text{F2}|V = /\text{i}/)$ is shifted down when the talker is male relative to when the talker is female.
- For another example, consider a talker with a very strong French accent, who produces word-initial voiced stops as prevoiced, and voiceless as unaspriated. This speaker's VOT distributions are shifted so far that their $p(\text{VOT}|p)$—with a mean of about 0—is the same as a native English speaker's $p(\text{VOT}|b)$.
- In order to maintain robust comprehension, listeners need to be sensitive to these differences in the distributions of cues.

3.1 Sensitivity to distributional information

- Clayards et al. (2008): listeners change their category boundary between /b/ and /p/ based on the *variance* of the VOT distributions associated with each: steeper slope for lower variance, as predicted by the ideal listener model.
- Munson (2011): similarly, listeners change their boundary based on the mean VOT values for /b/ and /p/ that they obesrve: boundary is shifted down when the VOT means are shifted down, and up for up.

3.2 Incremental belief updating

- How do poeple achieve this sensitivity to distributional information??
- The problem is that the listener doesn't have access to the true distribution (likelihood)
- They have to work with uncertain *beliefs* about the likelihood function.
- We can think of these beliefs as a probability distribution over category parameters (e.g. mean and variance of the category's cue distribution), $p(\mu_c, \sigma_c^2)$.
- This leads naturally to the idea of *belief updating*, where prior beliefs are brought into better alignment with recent experience by another application of Bayes Rule:

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

- That is, we can think of adaptation as an *inference* process: in order to deal with changes in the likelihood function (cue distribution), the listener has to infer the category means and variances based on the observed cue values (and other sources of information that might be available to disambiguate).
- This analysis makes two qualitative predictions: first, adaptation should depend on the statistics of the linguistic stimuli that are being adapted to, and second, adaptation should depend on the listener's prior beliefs about how the adapted categories are likely to be realized acoustically. This applies both to what kind of adaptation should occur (based on the mean and variance of the exposure stimuli, and which means and variances are most credible a priori), and how much adaptation should occur (based on the number of adapting stimuli encountered, and the confidence or "effective sample size" of the prior beliefs).

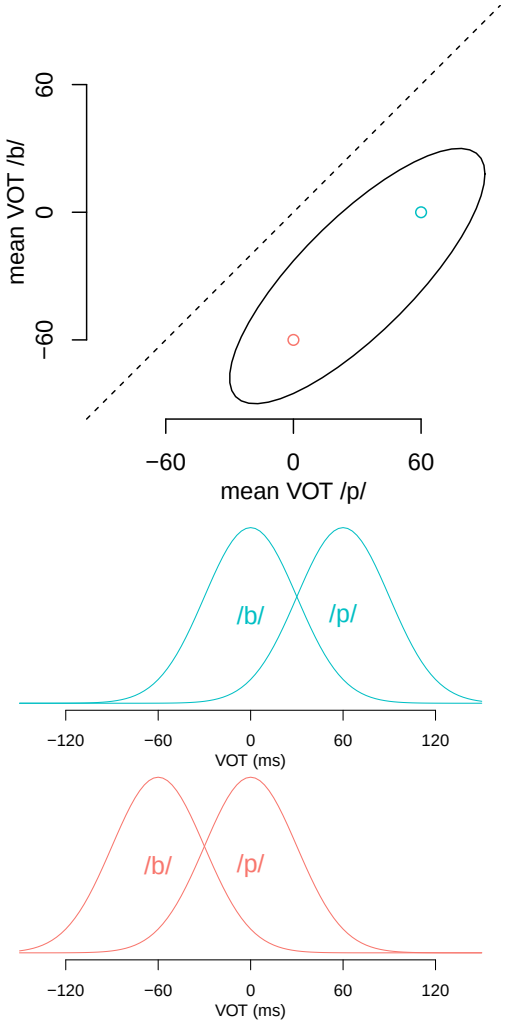
3.3 Modeling phonetic recalibration

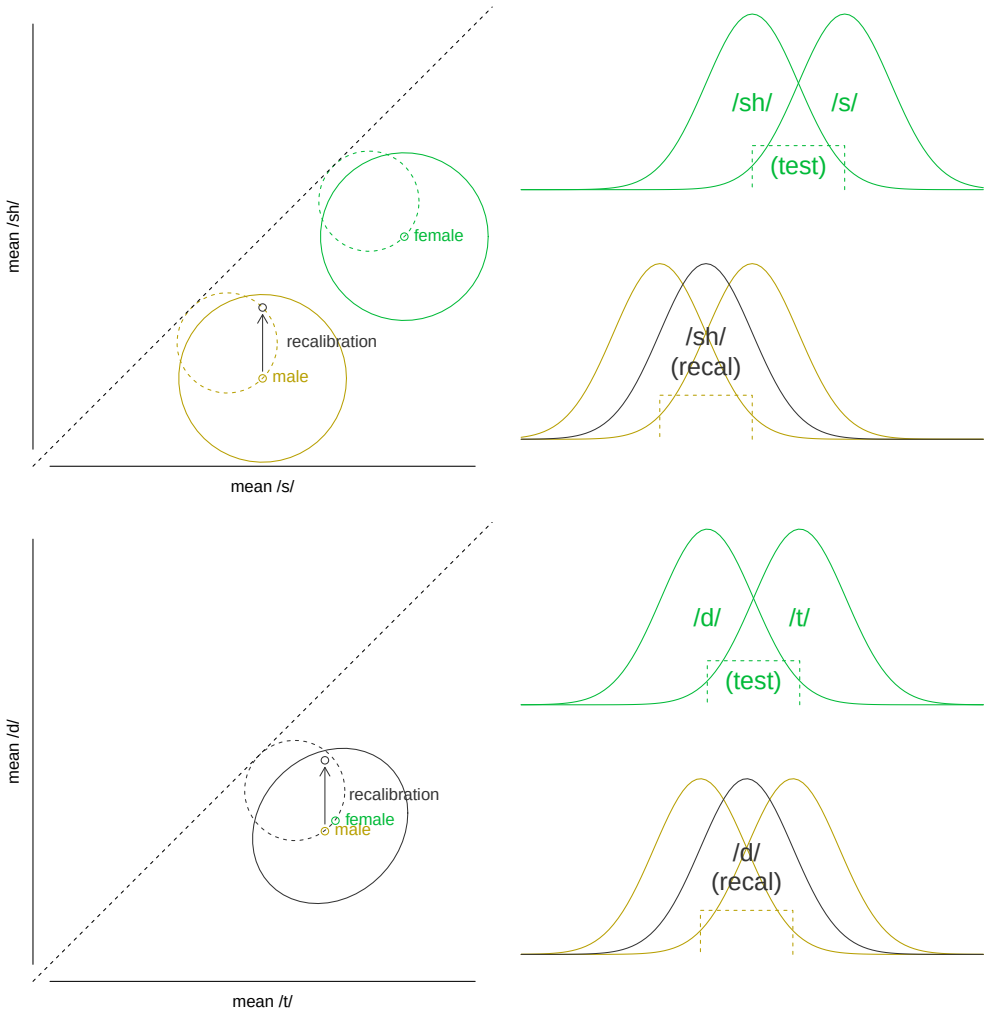
4 When to adapt?

- As a framework for thinking about how and when to adapt, it is useful to think of the true likelihood function in a given situation as a single point in a high dimensional *parameter space*, where each dimension is one category parameter (e.g. mean VOT value of /b/).
- Recall that the listener's beliefs are distributions over these parameters, and hence can be thought of as distribution in this space. The variance of this distribution corresponds to *uncertainty* on the part of the listener about the likelihood.
- The listener's prior beliefs are also a distribution in this space.
- To simplify, let's just consider the mean VOT value for /b/ and /p/. (Adding other cues, like vowel length, or other parameters, like the variance, and other categories, like /d/, just increases the dimensionality of this space)

- This representation is nice because it provides an easy way to represent prior beliefs about relationships between the cues associated with different categories like, say, the ordering of two categories' mean cue values, even if the actuall mean values themselves vary.

  We can use a positive correlation between the means to represent the fact that, the mean VOT value for /p/ is generally larger than for /b/, even though the actual means themselves vary quite a bit across situations:



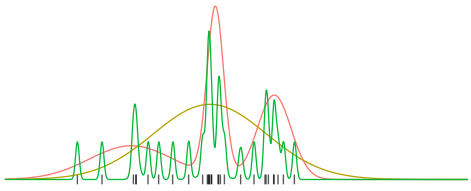


- Speech perception requires good—but not perfect—beliefs about the right likelihood to use. The "true" likelihood function is characterized by a single point in this space, and success is achieved when the beliefs are peaked around this point.
- Uncertainty—the beliefs not being peaked enough—and error—being peaked around the wrong value—can both be problematic.
- The listener's prior expectations are very important. If the prior beliefs are already (somehow) slightly peaked in the neighborhood of the true parameters, then only a small amount of experience is required for the listener to achieve accurate beliefs.
- If on the one hand the prior is too broad, then it takes more evidence to narrow things down enough.
- If, on the other, the prior is too peaked but not around the true value, then this leads to *errors*, where the listener's beliefs are pulled away from the true value, until enough evidence has accumulated that the prior may be overcome.
- Prior beliefs $p(\mu_c, \sigma_c^2)$ determine how much—and at the extreme, when—to adapt.

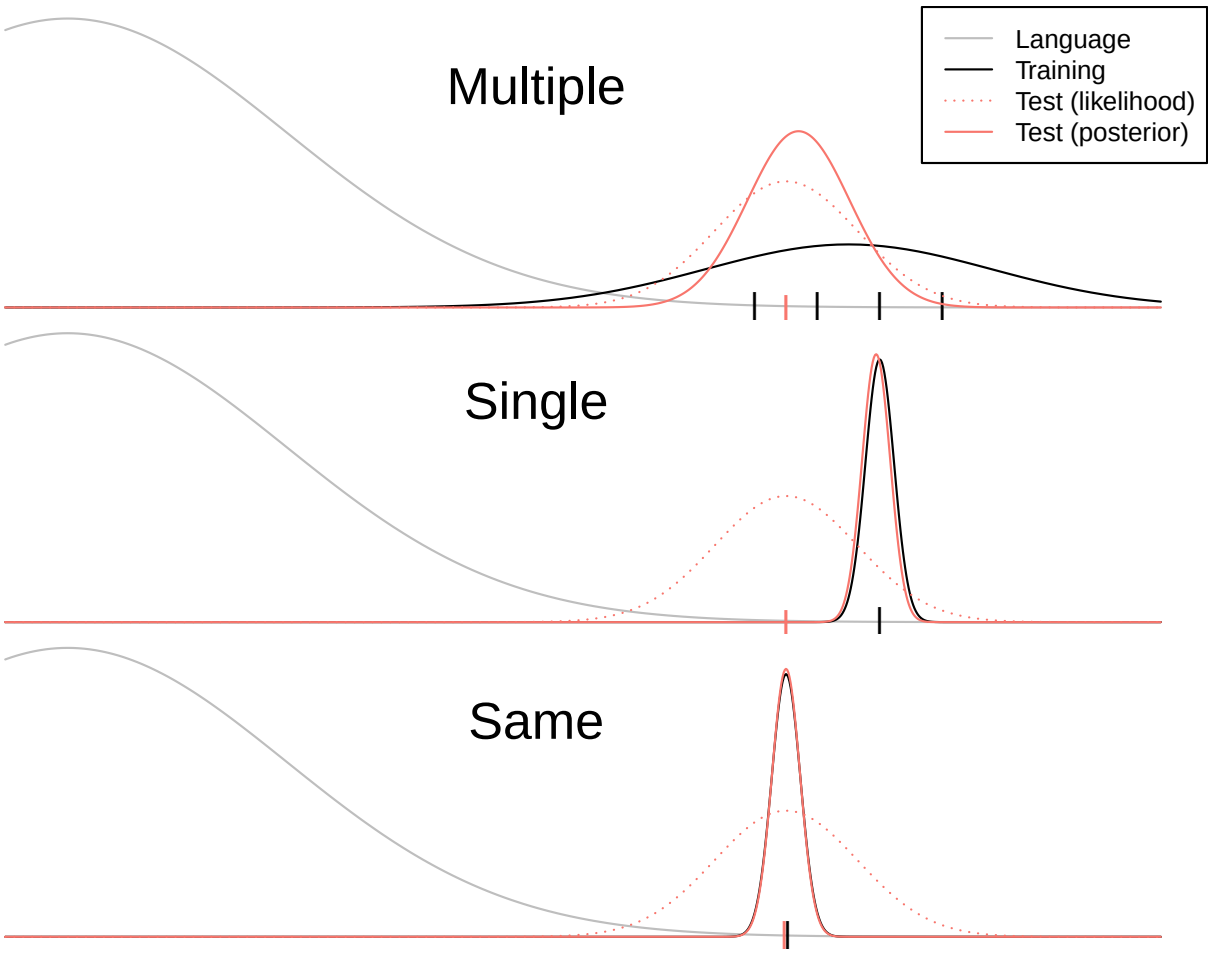
- A weak prior is required to produce the kind of rapid adaptation that is commonly observed in phonetic recalibration, which suggests that the effective prior sample size is very low.
- But the *true* sample size is very high: we have lots of experience with how phonetic categories are realized acoustically, having heard at least millions of productions of each phonetic category. Correspondingly, adaptation to a single, odd talker doesn't mess up our ability to comprehend other talkers.
- Why? This appears to be a paradox for this framework.
- The problem comes from (tacitly) assuming that the prior is the *same* for every situation. This only makes sense (in a statistical sense) when we assume that the changes in the category parameters (μ_c, σ_c^2) are just random fluctuations. In this case, the right thing to do is just keep your prior weak enough to track the changes as they come.
- But in reality, the changes in category distributions are systematic and structured. At the most basic level, they depend on the *who* is talking, or more generally, the context or environment.
- This leads, intuitively, to the notion of a *hierarchical* model for phonetic categories, where beliefs about phonetic categories are conditioned on speaker s (or context, more generally), $p(\mu_c, \sigma_c^2|s)$.

4.1 How lumpy?

- This means that a listener's overall prior beliefs about what category parameters are expected is a mixture prior, or a "lumpy prior", with one lump for each speaker $p(\mu_c, \sigma_c^2) = \sum_s p(\mu_c, \sigma_c^2|s)p(s)$
- Having one lump for all speakers and having one lump for *each* speaker two extreme ways of grouping speakers into clusters. To be optimal, the prior over parameters should reflect the degree of lumpiness in speakers that the listener has experienced, which is probably somewhere between.



4.2 Generalizing experience with foreign accented speech



- Bradlow and Bent (2008): experience with foreign accented talker improves comprehension of that talker.
- Experience with four other talkers, from the same native-language background, improves comprehension just as well for the first talker.
- But an equivalent amount of experience with just one of these talkers produces no gains in comprehension.