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

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

• Good comprehension depends on accurate likelihood p(x|c, μc, σc) (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.

PERCEPTUAL MAGNET EFFECT

• Influence of categories pulls percept towards category mean

• Inter-speaker’s intended target cue value based on observed cue value and knowledge of distributions (category variance and noise)

COMPENSATION FOR COARTICULATION

• Vocal-to-vowel coarticulation: first vowel takes on characteristics of following vowel.

• Listeners compensate for this by shifting their category boundaries.

• Model of conditioning likelihood for first vowel on second vowel

PROCEDURE

• Measure categorization responses to first vowel in 2V/3V2 words (V1 is /a/ or /i/; 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, μc, σc) (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

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

• 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.


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

ADAPTED CATEGORY BOUNDARY DUE TO VARIANCE CHANGES

• Category boundary slope reflects uncertainty in classification

• Steeper for lower variance distributions

• Listeners exposed to new 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

• Speakers are characterized by the parameters of their category likelihoods p(x|c, μc, σc)

• Prior beliefs about category parameters p(μc, σc) 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: fast. Prior is weak, must re-adapt every time environment changes.

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

Structured variation due to speakers: 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 al. 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 env.

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 (/t/-/d/) or fricative place (/b/-/d/) 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, 2006)

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 different Mandarin-accented talkers (one quarter as much on each) Results: Same and multiple talker training both produced 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).