Modeling cue re-weighting in dimension-based statistical learning

Yiming Lu (yimil40@uci.edu)
Department of Language Science, University of California, Irvine CA
Xin Xie (xxie14@uci.edu)
Department of Language Science, University of California, Irvine CA

Abstract

Speech perception requires inferring category membership from varied acoustic cues, with listeners adeptly adjusting cue utilization upon encountering novel speech inputs. This adaptivity has been examined through the dimension-based statistical learning (DBSL) paradigm, which reveals that listeners can quickly de-emphasize secondary cues when cue correlations deviate from long-term expectations, resulting in cue-reweighting. Although multiple accounts of cue-reweighting have been proposed, direct comparisons of these accounts against human perceptual data are scarce. This study evaluates three computational models–cue normalization, Bayesian ideal adaptor, and error-driven learning–against classic DBSL findings to elucidate how cue reweighting supports adaptation to new speech patterns. These models differ in how they map cues onto categories for categorization and in how recent exposure to atypical input patterns influences this mapping. Our results show that both the error-driven learning and ideal adaptor models effectively capture the key patterns of cue-reweighting phenomena, whereas prelinguistic cue normalization does not. This comparison not only highlights the models' relative efficacy but also advances our understanding of the dynamic processes underlying speech perception adaptation.

Keywords: Dimension-based statistical learning; Error-driven learning; Ideal adaptor model; Speech perception; Adaptation

Introduction

Speech perception, like many other types of perceptual categorization, requires integrating multiple sources of information in a high-dimensional space. For instance, listeners may use up to 16 distinct acoustic-phonetic cues to differentiate the phonemes /b/ and /p/ in words like 'bear' and 'pear' (Lisker, 1986). Importantly, native listeners are sensitive to the statistical distributions of these cues in spoken language and assign variable 'perceptual weights' to them. In English, voiced stops (/b/, /d/, /g/) are mostly saliently contrasted with voiceless stops (/p/, /t/, /k/) in production along two acoustic dimensions: Voice Onset Time (i.e., the timing differences between the onset of voicing relative to the release of a stop, henceforth VOT) and fundamental frequency (f0) at vowel onset. As shown in Fig. 1, VOT is generally more diagnostic for voicing than f0. Indeed, previous research shows that native-English listeners use VOT as a primary cue and f0 as a secondary cue when discerning voicing distinctions (Dmitrieva et al., 2015; Whalen et al., 1993).

The relative weighting of cues, however, is not fixed, but rather malleable. Cue re-weighting happens during first language acquisition. Children typically do not rely on f0 in perceiving the voicing distinction until at least age 6 (Bernstein, 1983). By some estimates, the increasing reliance on f0 continues into late adulthood (Toscano & Lansing, 2019), suggesting a prolonged learning process for cue weights. Adjustment in cue weighting is also important for L2 sound category learning (Francis & Nusbaum, 2002). More recently, cue-reweighting processes have been proposed as a key mechanism underlying adaptation to talker accents (Idemaru & Holt, 2011; Xie et al., 2017; Wu & Holt, 2022), operating over much shorter timescales.

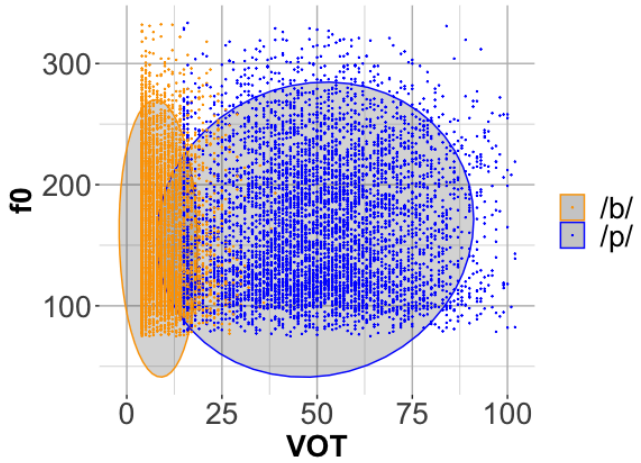

Figure 1: The distribution of raw VOT and f0 for /b/ and /p/ in Chodroff & Wilson (2018), a large-scale speech corpus for word-initial prevocalic stops in American English. Each dot represents an instance of speech tokens. Yellow dots represent /b/, while blue dots represent /p/. Ellipses show bivariate Gaussian 95% CI of each category.

One of the clearest demonstrations of cue-reweighting as a result of short-term exposure comes from the Dimension-based Statistical Learning paradigm (DBSL, henceforth; Idemaru and Holt, 2011, 2014, 2020; Hodson et al., 2023; Lehet & Holt, 2020; Liu & Holt, 2015; Schertz et al., 2016; Zhang & Holt, 2018; Zhang et al., 2021). In DBSL, participants are exposed to cue distributions that deviate from their long-term experience. For instance, in English, VOT and f0 tend to be positively correlated across stop categories such that both longer VOTs and higher f0s signal voicing (e.g., /b/ vs. /p/). When listeners are exposed to input distributions where the correlation between VOT and f0 is reversed, listeners quickly down-weight their reliance on f0 to the extent that variation in f0 no longer affects their voicing judgments in categorization tasks. Such cue reweighting effects have been observed for other types

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

of segmental and suprasegmental contrasts (Liu & Holt, 2015; Lehet & Holt, 2020; Wu & Holt, 2022).

Beyond the cue-reweighting processes as observed in DBSL, multiple lines of research using other paradigms have also revealed rich evidence of short-term adaptation as a way for listeners to cope with acoustic variability in speech (Norris et al., 2003; Samuel & Kraljic, 2009). While it is commonly believed that adaptation occurs when the acoustic-phonetic regularities of the exposure input (e.g., a negative VOT-f0 correlation) deviate from listeners' long-term knowledge (e.g., shorter VOTs signal voicing), the exact mechanisms underlying this adaptation remain widely debated (Harmon et al., 2019; Kleinschmidt & Jaeger, 2015; Xie et al., 2023).

Some theories propose that listeners encode the cue-to-category relationship for each cue independently, assigning weights to each cue based on its reliability in signaling phonetic contrasts (Toscano & McMurray, 2010; Harmon et al., 2019). Categorization then reflects a weighted sum of these cue values. How are these weights updated? One mechanism is error-driven learning[1]. When listeners hear a speech token, they categorize it based on previously learned connection weights between cues and category labels. If the acoustic input violates expectations–such as encountering speech patterns with unexpected category labels–prediction error occurs, serving as a feedback signal for adjusting weights (Hodson et al., 2023). Although often evoked as a potential learning mechanism for cue-reweighting, the specifics of error-driven learning are rarely formalized (but see Harmon et al., 2019).

Another approach posits that categories are organized in a multidimensional space; listeners not only encode category-specific distributions along each cue dimension but also track the covariation among multiple cues (Kleinschmidt & Jaeger, 2015; Xie et al., 2023). In these models, cue weights are not explicitly represented; instead, they emerge as listeners become sensitive to cue variance and covariance. Crucially, updating category parameters –such as category means and variance-covariance–relies on the integration of prior beliefs (estimates based on long-term input statistics) and current data (estimates from the observations during exposure). This belief updating process has been formalized in Bayesian ideal adapter models, which have been shown to effectively account for various phenomena in adaptive speech perception (Hitczenko & Feldman, 2016; Xie et al., 2023).

Yet another competing view, although not mutually exclusive with the previous two, holds that short-term adaptation may be driven by computationally simpler mechanisms that do not require changes in category representations. Specifically, raw acoustic cues are 'normalized' pre-linguistically prior to being mapped onto categories to accommodate variability in the input (Jongman & McMurray, 2009; Barreda, 2020; Persson & Jaeger, 2023). As such, change in normalization schemes, rather than category representations, may be sufficient to explain adaptive changes in perception. For instance, normalizing a token's specific VOT relative to a talker's mean VOT values helps to reduce cross-talker differences in category realizations, leading to less overlap between contrastive categories in the normalized cue space relative to raw cue space (Jongman & McMurray, 2009). Although traditionally deemed incapable of accounting for adjustments to complex alterations in the acoustic input, recent work has demonstrated that exposure-elicited updating of cue normalization schemes can qualitatively explain adaptation to natural nonnative accents (Xie et al., 2023).

In sum, although cue reweighting has been robustly observed in human perceptual experiments, the mechanisms supporting it remain poorly understood. In this paper, we evaluate three computational models, each representing a different theoretical approach to adaptation. We revisit the classic experiment of Idemaru and Holt (2011) and simulate human responses using three models: error-driven learning, Bayesian ideal adaptor, and cue normalization change model. Aside from the error-driven learning model, which was implemented by Harmon et al., (2019), these models have not previously been applied to the DBSL paradigm. Extending from Harmon et al. (2019), we use input statistics from a naturalistic speech corpus (Chodroff & Wilson, 2018) to estimate listeners' long-term knowledge. Our findings indicate that both error-driven learning and Bayesian ideal adaptor models offer a qualitative fit to the empirical data, despite their differing assumptions about category structure and learning processes.

In what follows, we first describe the experiment from Idemaru & Holt (2011) and then present the simulation results.

Idemaru and Holt (2011)

Idemaru and Holt (2011) (IH11, henceforth) investigated how short-term exposure to speech regularities that deviate from listeners' long-term experience changes cue weighting. In experiment 1, they examined the use of f0 and VOT in voicing distinction for two contrasts (/d/-/t/, /b/-/p/). Here we focus on the bilabial stops only. Specifically, participants heard synthesized speech tokens in three exposure blocks (canonical, neutral, and reversed), presented in a sequential order. Each exposure block contained 10 iterations; each iteration contained 10 exposure trials and 2 test trials, which were interleaved and presented in a random order. The three exposure blocks were created by changing the pairing between VOT (short vs. long) and f0 (high vs. low), thereby changing the cue correlations. In the canonical block, short VOTs were paired with low f0s, whereas long VOTs were paired with high f0s, resulting in a positive VOT-f0 correlation, which mirrors the statistical regularities in American English. In the neutral condition, both short and

[1] Note that error-driven learning is not conceptually bound to any particular categorization model. It can be implemented with both cue integration and multidimensional models.

long VOTs were paired with mid-range f0s, making f0 uninformative in cueing voicing. In the reversed condition, low VOTs were paired with high f0s, resulting in a negative VOT-f0 correlation, which reversed the pattern observed in American English. Test items had ambiguous VOT values such that participants could only rely on f0 (high vs. low) to make voicing judgments. Throughout the experiment, participants categorized speech tokens (e.g., /b/ or /p/) without feedback. The results of the experiment are shown in Figure 2. The proportions of /p/ responses were significantly different for high and low f0 tokens in the canonical and the neutral blocks, suggesting that listeners relied on f0 to make voicing distinctions. However, this difference diminished in the reversed block. This indicates that after hearing the reversed cue correlations, listeners did not rely on varying f0 in making the voicing distinction to the same extent, signaling a down-weighting of f0.

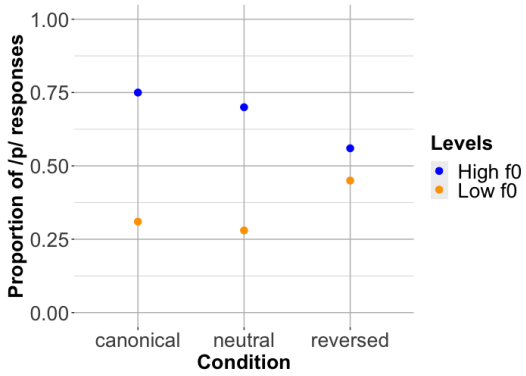

Figure 2. Results of Idemaru & Holt (2011)'s experiment 1. The x-axis shows the three exposure blocks and the y-axis shows the proportion of /p/ responses. The level of f0 values were color-coded with blue dots representing the high f0 (290Hz) and yellow dots representing low f0 (230Hz).

A few aspects of the experiment are noteworthy. First, despite no explicit category labels being provided during categorization, VOTs continued to serve as an informative cue for category membership. The authors reasoned that "listeners would recognize the 'accented' words during the robust VOT cue" while tracking how f0 covaried with VOT. This hypothesis was bolstered by recent evidence that the informativity of the primary cue (e.g., VOT) affects the magnitude of the down-weighting of the secondary cue (e.g., f0) when the two cues offer conflicting information as in the reversed block (Wu & Holt, 2022). In all simulations, we build on this hypothesis and assume that category labels are available to listeners in each trial.

Second, although the three blocks were designed to contain distinct acoustic regularities, they were sequentially presented to participants without breaks. Therefore, the down-weighting effect observed in the reversed block may reflect carryover effects from the previous exposure blocks (canonical and neutral). In a follow-up experiment, f0 down-weighting was replicated after listeners heard consecutive reversed blocks (Idemaru & Holt, 2011). However, it did not rule out the possibility that listeners may cumulatively track the distributions of speech input. This idea is well supported by evidence from talker-specific accent adaptation using the lexically-guided perceptual recalibration paradigm (e.g., Tzeng, Nygaard, & Theodore, 2021; Saltzman & Myers, 2021) and has been formally modeled in ideal adaptor models via incremental belief updating (Kleinschmidt & Jaeger, 2015; Xie et al., 2023). On the other hand, it is also possible that upon detecting (whether explicitly or implicitly) changes in the input statistics, listeners reset their cue weights to those learned from long-term experience from block to block (Hodson, Shinn-Cunningham, & Holt, 2023). In our simulations, we consider both possibilities and compare the model performance in two scenarios: *independent* use of block-wise cue distributions vs. *sequential* use (i.e., cue weights updated according to the cumulative distributions across blocks).

Lastly, although the reliance on f0 was down-weighted in the reversed block, the f0-to-category mapping was never reversed. That is, listeners did not report hearing /p/ in the presence of low f0 values. One possibility is that while listeners tuned in to local perturbations, their long-term knowledge constrained the short-term adaptation. We varied the parameters controlling the learning speed to simulate how the relative weighting of prior experience and short-term exposure affects cue-reweighting.

Simulation of cue-reweighting in DBSL

We take the following steps to simulate Experiment 1 from IH11. First, we estimate listeners' prior beliefs assuming that their category knowledge corresponds to the long-term distributional properties in natural productions. Second, we update the categories based on the acoustic distributions participants heard during exposure. Lastly, we predict the proportion of /p/ responses for the test tokens based on each model's updated categories.

We investigate whether computational models can replicate the qualitative response patterns observed for the test stimuli across three blocks. In particular, we examine whether f0 is down-weighted in the reversed block, as indicated by decreased differentiation in categorization responses between tokens with high versus low f0 values. In addition, we assess whether, across blocks, these response differences retain the long-term f0/VOT correlations, where high f0 led to more /p/ responses than low f0, mirroring human response patterns.

Estimating long-term knowledge

Long-term knowledge of categories is estimated from a speech corpus of 180 native English speakers (Chodroff and Wilson, 2018). Here, we used VOT and f0 (Mel) for the /b/-/p/ contrast only, and used a balanced sample from each talker ($N \geq 33$ tokens per category per talker). A total of 11124 tokens from 96 talkers were used to estimate prior category knowledge.

Exposure and test data

Following IH11, exposure items contained three levels of VOTs typical for each category: -20, -10, and 0ms VOTs for /b/, and 20, 30, and 40ms for /p/. For the canonical trials, the VOTs for /b/ were paired with low f0s (220, 230, 240Hz), and /p/ with high f0s (280, 290, 300Hz). For the neutral trials, both /p/ and /b/ VOT values were paired with mid-range f0s (250, 260, and 270Hz). For the reversed trials, the VOTs for /p/ were combined with low f0s (220, 230, and 240Hz), and /b/ with high f0s (280, 290, and 300Hz). The test items were created by pairing 10ms VOT with f0 values (230Hz and 290Hz for low and high f0s, respectively). Each exposure item was assigned a category label (/b/ or /p/) based on its VOT values. These labels were used in all models below.

Preprocessing data

We converted raw cues into talker-normalized cues by employing a general normalization scheme (McMurray & Jongman, 2011) that centers cues relative to a talker's cue mean. This procedure was performed on all data used to estimate long-term knowledge as well as the exposure and test tokens used in IH11. Exposure and test tokens were treated as tokens from a single talker in normalization.

Change in category representation

We consider two separate possibilities by which adaptive changes can occur as a function of exposure: 1) changes in category representations (i.e., cue-to-category mapping) and 2) changes in cue normalization. Under 1), we consider two distinct mechanisms by which short-term exposure to various cue distributions drives changes in listeners' category representations: error-driven learning and Bayesian belief updating as implemented in ideal adaptor models.

Error-driven learning We adopted the error-driven learning model from Harmon et al. (2019). This model updates cue weights upon encountering prediction errors. Firstly, the activation weights (w_t) for each possible category at the current state (t) is calculated by summing the product of cue (C) values and their weights (w) (Eq.1). The activation weights for both categories are calculated. Secondly, if a category is observed, the cue weight will be incremented by the product of 1-$\sum w_t$ and the learning rate r (Eq.2). If a category is incorrectly categorized, the cue weights will be incremented by the product of 0-$\sum w_t$ and the learning rate r (Eq. 3). Assuming Luce's choice rule, we take the activation weights of /p/ for test items as the proportion of /p/ responses.

$$\sum w_t = w_{t,VOT} * C_{VOT} + w_{t,f0} * C_{f0} \quad (1)$$

$$w_{t+1,C} = w_{t,C} + (1 - \sum w_t) * r \quad (2)$$

$$w_{t+1,C} = w_{t,C} + (0 - \sum w_t) * r \quad (3)$$

Ideal adaptor models. We adopted the ideal adaptor model from Xie et al., (2023), which extended Kleinschmidt & Jaeger (2015) to multivariate acoustic inputs. This model assumes that listeners form generative models based on their prior experience, while maintaining uncertainty about their beliefs. The beliefs can be updated based on recent input. Specifically, categories correspond to bivariate Gaussian distributions of multiple cues, parameterized by their cue means ($\boldsymbol{\mu}_{t,C}$) and covariance matrices ($\boldsymbol{\Sigma}_{t,C}$). When encountering a new talker, listeners estimate category mean ($\bar{x}_C$) and covariance ($S_{x,C}$) from the exposure tokens. With an increasing number (N_C) of tokens, listeners update the mean and covariance matrices for each category according to Eq. 4. Two parameters $\mathbf{K}_C$ and $\boldsymbol{\nu}_C$ control the relative weight of prior beliefs compared to the exposure input received from the new talker in the updating process, with larger values simulating slower learning of changes in the category mean and covariances respectively. For instance, given 100 trials of exposure, a $\mathbf{K}_C$ = 100 would be that listeners' estimate of the category mean would be equally affected by their prior knowledge of the category mean and the estimated mean of exposure stimuli. For simplicity, we assume that the $\mathbf{K}_C$ and $\boldsymbol{\nu}_C$ are identical for /b/ and /p/.

$$\boldsymbol{\mu}_{t+1,C} = \frac{1}{K_C+N_C}(\mathbf{K}_C\boldsymbol{\mu}_{t,C} + N_C\bar{x}_C) \quad (4)$$

$$\boldsymbol{\Sigma}_{t+1,C} = \frac{\boldsymbol{\nu}_C}{\boldsymbol{\nu}_C+N_C}\boldsymbol{\Sigma}_{t,C} + \frac{N_C}{\boldsymbol{\nu}_C+N_C}(S_{x,C} + \frac{\mathbf{K}_C}{\mathbf{K}_C+N_C}(\bar{x}_C - \boldsymbol{\mu}_{t,C})(\bar{x}_C - \boldsymbol{\mu}_{t,C})^T)$$

Changes in pre-linguistic cue normalization

The normalization change model implements the process by which listeners update their estimate of cue mean. Before exposure, listeners rely on the cue mean expected from one's long-term experience (i.e., the population mean). With increasing exposure, listeners update the estimated talker-specific cue mean based on the recent input. This model estimates the prior talker mean (μ_t) from long-term input statistics with some degree of confidence (**K**). The listener incrementally updates the belief of cue mean (μ_{t+1}) with a growing number of tokens (N) from the new talker's input ($\bar{x}$) according to Eq.5.

$$\mu_{t+1} = \frac{1}{\mathbf{K}+N}(\mathbf{K}\mu_t + N\bar{x}) \quad (5)$$

Simulation results

Next we present the simulation results and compare them to the human responses in IH11. For each model, we consider the possibility that listeners treat each block separately (*independent* model) or cumulatively (*sequential* model).

Independent models are trained on a specific exposure block alone, while sequential models take the input from previous blocks into account. Specifically, we used the updated parameters (e.g., $\boldsymbol{\mu}_{t,C}, \boldsymbol{\Sigma}_{t,C}, \mathbf{K}_C, \boldsymbol{\nu}_C$) after previous blocks as priors for subsequent blocks.

Error-driven learning

Figure 3 shows the predictions from the error-driven learning models. First, consider the independent models (Fig.3 left). Mirroring human responses, low f0 was predicted to induce fewer /p/ responses than high f0 in the canonical block. For the neutral block, no difference was predicted between the two f0 levels. This indicates a down-weighting of f0 not observed in human responses where the distinction between the two f0 levels remained salient as in the canonical block. For the reversed block, the model predicted an inverted pattern: a high f0 token became strongly predictive of /b/, whereas a low f0 indicated /p/. This pattern also deviates from IH11, where the f0-to-category mapping was never reversed.

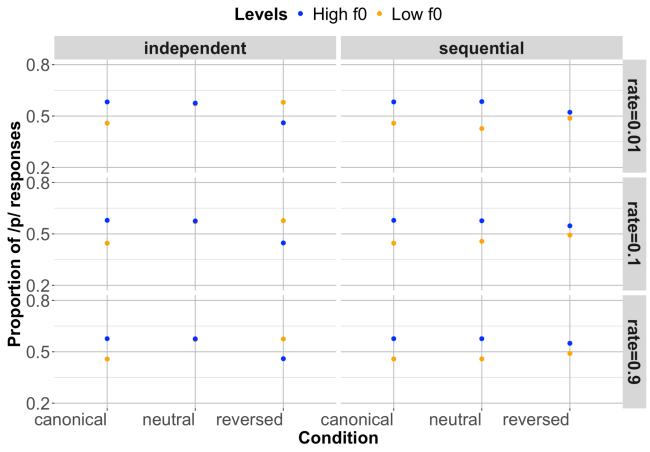

Figure 3: Predicted categorization performance of an error-driven learning model, assuming independent (left) vs. sequential (right) tracking of input regularities presented across blocks. The learning rate r is set to 0.01, 0.1, and 0.9, respectively.

Unlike the independent models, the sequential models (Fig.3 right) qualitatively reproduced the human response pattern in all three blocks, where down-weighting only occurred in the reversed block. The direct contrast between the independent and the sequential models is compatible with the idea that listeners cumulatively track input statistics spanning multiple blocks and update their category representations accordingly. Of note,, the directionality of cue reweighting was highly consistent across different learning rates.

Ideal adaptor model

To evaluate the impact of belief-updating speed on cue reweighting, we compared different levels of strength of the prior beliefs by varying **K** and **ν**. Since the effect of **ν** turns out to be relatively small, here we only illustrate the different results for **K**, with **ν** held constant at 256 (i.e., listeners primarily rely on their prior knowledge of cue covariance).

Figure 4 (left) shows the simulation results for the independent models. First of all, consistent with human responses, predictions for the canonical block successfully captured the human pattern. Moving to the neutral block and the reversed block, the model predicted a down-weight of f0–similar to the independent error-driven learning models, although the magnitude of downweighting was clearly mediated by the speed of belief updating in the ideal adaptor models. When the updating of category mean was slow (i.e., **K** was large), the independent model qualitatively capture the human responses, yielding a clear downweighting only in the reversed block. At faster learning rates, the model reversed f0-to-category mapping, which did not occur for human listeners.

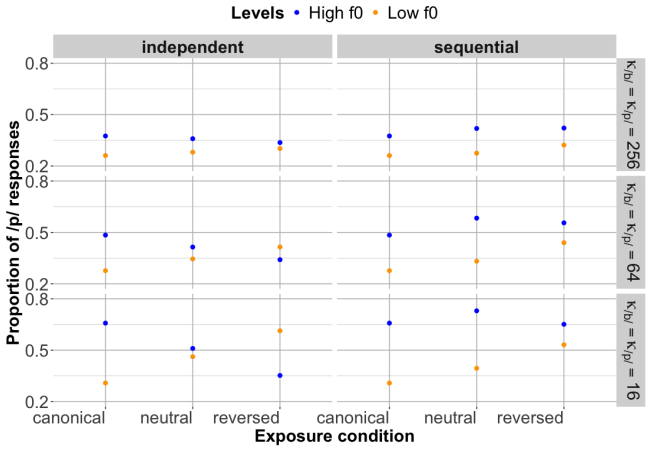

Figure 4. Predicted categorization performance for an ideal adaptor model, assuming independent (left) vs. sequential (right) tracking of input regularities presented across blocks. This figure shows the results when the strengths of the prior belief of the category mean (**K**) are set at 16, 64, and 256, and the strength of the prior belief for the covariance matrix (**ν**) is set at 256.

Figure 4 (right) presents the simulation results for the sequential ideal adaptor models. Different from the independent models, the human response patterns were qualitatively replicated throughout all three blocks at all parameterization in that f0-to-category mapping was not reversed in the reversed block in any parameterization. In addition, the difference in the proportion of /p/ responses between high and low f0 was larger in the canonical and neutral blocks, and smaller in the reversed block, signaling a down-weighting of f0. The magnitude of down-weighting was similar to that observed in the human responses when learning was fast (i.e., **K** was small).

Cue normalization change models

The change models of cue normalization yielded the same results under the independent and sequential assumptions (Fig. 5). The proportion of /p/ responses was identical across the three blocks, suggesting that none of the models succeeded in simulating the down-weighting effects.

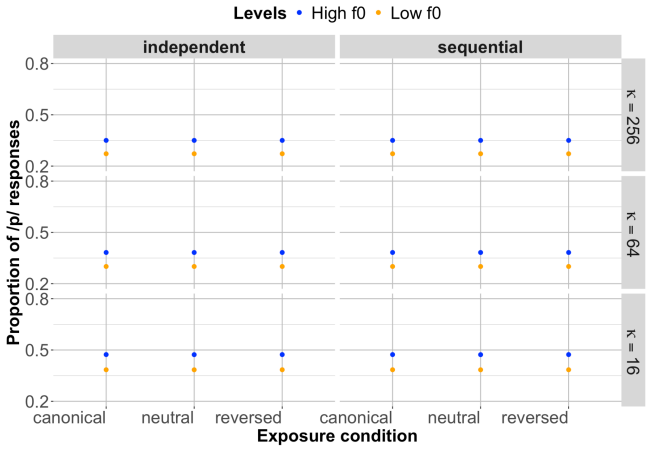

Figure 5: Predicted categorization performance for a normalization model, assuming independent (left) vs. sequential (right) tracking of input regularities presented across blocks. Here, we present the proportion of /p/ response when the strengths of the prior belief of talker mean (**K**) is set at 16, 64, and 256, respectively. The two models yield identical results.

General Discussion

Evidence from dimension-based statistical learning indicates that listeners can rapidly adjust the perceptual weights of specific acoustic cues as a function of recent exposure (Idemaru & Holt, 2011; Liu & Holt, 2015). What learning mechanism(s) support cue-reweighting? To answer this question, we simulated the IH11 experiment using three computational models. These models implement different theoretical accounts of how cues are mapped onto categories as well as whether and how the cue-to-category mappings change as a function of recent input. Each model was trained on the exposure tokens from the canonical, neutral, and reversed blocks as used by the original IH11 study. For each model, we compared its performance under two alternative assumptions: 1) listeners independently track cue distributions within each block and 2) cumulatively track cue distribution across blocks.

Under the independent assumption, listeners do not carry over the adaptation to the input from one block to another. The independent error-driven model failed to capture the IH11 results. The independent ideal adaptor model could capture the qualitative trend at a slow learning rate. Yet, it significantly fell short of simulating the numerical differences in the proportion of /p/ responses. In contrast, when cumulative learning across blocks was assumed, both models qualitatively captured the trajectories of adaptive perception. Beyond DBSL, evidence that listeners track input cumulatively is also available from lexically guided perceptual learning (Tzeng, Nygaard, & Theodore, 2021; Saltzman & Myers, 2021), where previous exemplars of a novel talker can be used further downstream.

Secondly, comparing the three classes of models, it is clear that changes in pre-linguistic cue normalization do not elicit cue re-weighting regardless of the parameterization. This is because changes in cue normalization are driven by a change in cue mean, yet in the specific experiment simulated here, f0 mean was intentionally kept constant across all three blocks. In contrast, error-driven learning and ideal adaptors can both qualitatively capture human responses in IH11, despite the fundamental differences in their architecture. Noticeably, the error-driven models predict smaller differences between tokens with high versus low f0s than was observed among human listeners, regardless of learning rates. This inadequacy could stem from the Rescorla-Wagner learning rule and/or the cue integration during categorization. More simulations are underway to probe into the reason further.

Interestingly, the absolute differences in humans' /p/ responses between high and low f0 tokens were best approximated by ideal adaptors at faster learning rates (e.g., **K** = 16). Our results therefore imply that human participants in IH11 were ready to discard their long-term knowledge within just a few trials. This finding is compatible with other evidence that demonstrates the rapidity of cue weighting in other DBSL studies. For instance, more recent studies have revealed that not only does the down-weighting of secondary cues occur rapidly, it also diminishes quickly with a few additional canonical trials (Idemaru & Holt, 2011; Liu & Holt, 2015; Zhang & Holt, 2018; Hodson et al., 2023).

Beyond short-term talker accent adaptation, cue reweighting has been proposed to phonetic learning in second language acquisition (Francis & Nusbaum, 2002; Holt & Lotto, 2006; Iverson et al., 2003). Yet the rapidity of the lab-elicited cue reweighting seems to stand in direct contrast to real-world scenarios where cue reweighting, as often required for second language phonetic learning, takes months, if not years, to occur. For example, Japanese learners of English can experience tremendous difficulty up-weighting F3 in the perception and production of English /r/-/l/ distinction (Yamada & Tohkura, 1992). How come DBSL induces fast learning, yet cue re-weighting in L2 learning appears so slow? One possibility is that when highly artificial speech tokens are used, listeners more readily change cue weighting; whereas in an experiment where the speech input more closely resembles real-world scenarios, listeners would be more conservative in updating their cue weights and employ a slower learning rate. How task demands and speech input characteristics affect listeners' interpretation of an experiment and consequently the cue reweighting is an open question.

To sum up, we show that distinct learning mechanisms may qualitatively capture the cue reweighting pattern, but the ideal adaptor can more flexibly adjust cue distributions than the error-driven model.

References

Barreda, S. (2021). Perceptual validation of vowel normalization methods for variationist research. Language Variation and Change, 33(1), 27–53. https://doi.org/10.1017/S0954394521000016

Bernstein, L. E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. Journal of Phonetics, 11(4), 383–393. https://doi.org/10.1016/S0095-4470(19)30837-X

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. Journal of Phonetics, 61, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. Journal of Phonetics, 49, 77–95. https://doi.org/10.1016/j.wocn.2014.12.005

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. Journal of Experimental Psychology: Human Perception and Performance, 28(2), 349–366. https://doi.org/10.1037/0096-1523.28.2.349

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. Cognition, 189, 76–88. https://doi.org/10.1016/j.cognition.2019.03.011

Hitczenko, K., & Feldman, N. H. (2016, August). Modeling adaptation to a novel accent. In Proceedings of the Annual Conference of the Cognitive Science Society.

Hodson, A., DiNino, M., Shinn-Cunningham, B., & Holt, L. L. (2022). Dimension-Based Statistical Learning in Older Adults. Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44). https://escholarship.org/uc/item/6qn2q5rc

Hodson, A. J., Shinn-Cunningham, B. G., & Holt, L. L. (2023). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. Cognition, 238, 105473. https://doi.org/10.1016/j.cognition.2023.105473

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. The Journal of the Acoustical Society of America, 119(5), 3059-3071.

Idemaru, K., & Holt, L. L. (2011). Word Recognition Reflects Dimension-based Statistical Learning. Journal of Experimental Psychology. Human Perception and Performance, 37(6), 1939–1956. https://doi.org/10.1037/a0025641

Idemaru, K., & Holt, L. L. (2014). Specificity of Dimension-Based Statistical Learning in Word Recognition. Journal of Experimental Psychology. Human Perception and Performance, 40(3), 1009–1021. https://doi.org/10.1037/a0035269

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. Attention, Perception, & Psychophysics, 82(4), 1744–1762. https://doi.org/10.3758/s13414-019-01956-5

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition, 87(1), B47-B57.

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. https://doi.org/10.1037/a0038695

Lehet, M., & Holt, L. L. (2017). Dimension-Based Statistical Learning Affects Both Speech Perception and Production. Cognitive Science, 41(S4), 885–912. https://doi.org/10.1111/cogs.12413

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. Cognition, 202, 104328. https://doi.org/10.1016/j.cognition.2020.104328

Lisker, L. (1986). "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. Language and Speech, 29(1), 3–11. https://doi.org/10.1177/002383098602900102

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), 219–246. https://doi.org/10.1037/a0022325

Persson, A., & Jaeger, T. F. (2023). Evaluating normalization accounts against the dense vowel space of Central Swedish. Frontiers in Psychology, 14, 1165742. https://doi.org/10.3389/fpsyg.2023.1165742

Saltzman, D., & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. Psychonomic Bulletin & Review, 28(4), 1354–1364. https://doi.org/10.3758/s13423-021-01885-1

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. Attention, Perception, & Psychophysics, 78(1), 355–367. https://doi.org/10.3758/s13414-015-0987-1

Toscano, J. C., & Lansing, C. R. (2019). Age-Related Changes in Temporal and Spectral Cue Weights in Speech. Language and Speech, 62(1), 61–79. https://doi.org/10.1177/0023830917737112

Toscano, J. C., & McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. Cognitive Science, 31.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. Psychonomic Bulletin & Review, 28(3), 1003–1014. https://doi.org/10.3758/s13423-020-01840-6

Whalen, D. H., Abramson, A. S., Lisker, L., & mody, maria. (1993). FO gives voicing information even with unambiguous voice.

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. Journal of Experimental Psychology: Human Perception and Performance, 48(9), 913–925. https://doi.org/10.1037/xhp0001037

Xie, X., Kurumada, C., & Jaeger, T. F. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/hgdn4

Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. Journal of Experimental Psychology: Human Perception and Performance, 43(1), 206–217. https://doi.org/10.1037/xhp0000285

Yamada, R. A., & Tohkura, Y. I. (1992). The effects of experimental variables on the perception of American English/r/and/l/by Japanese listeners. *Perception & psychophysics*, *52*, 376-392.

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. Journal of Experimental Psychology: Human Perception and Performance, 44(11), 1760–1779. https://doi.org/10.1037/xhp0000569

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information. Cognitive Science, 45(3), e12947. https://doi.org/10.1111/cogs.12947