Supervised and unsupervised learning in phonetic adaptation

Dave F. Kleinschmidt[1], **Rajeev Raizada**[1], and **T. Florian Jaeger**[1,2,3]

{dkleinschmidt, raizada, fjeager} @ bcs.rochester.edu

[1]Department of Brain and Cognitive Sciences, [2]Department of Computer Science, and [3]Department of Linguistics,
University of Rochester, Rochester, NY, 14607 USA

Abstract

Speech perception requires ongoing perceptual category learning. Each talker speaks differently, and listeners need to learn each talker's particular acoustic cue distributions in order to comprehend speech robustly from multiple talkers. This phonetic adaptation is a *semi-supervised* learning problem, because sometimes a particular cue value occurs with information that *labels* the talker's intended category for the listener, but other times no such labels are available. Previous work has shown that adaptation can occur in both purely *supervised* (all labeled) and purely *unsupervised* (all unlabeled) settings, but the interaction between them has not been investigated. We compare unsupervised with (semi-) supervised phonetic adaptation and find, surprisingly, that adult listeners do *not* take advantage of labeling information to adapt more quickly or effectively, even though the labels affect their categorization. This suggests that, like language acquisition, phonetic adaptation in adults is dominated by unsupervised, distributional learning.
Keywords: Cognitive Science, Linguistics, Psychology, Language understanding, Learning, Speech recognition

Introduction

Everyone speaks differently. In order to deal with this variability, listeners need to adapt to each new talker they meet, learning how they produce each phonetic category. For instance, in order to tell whether a talker intended to produce a /b/ or /p/, a listener needs to first learn that talker's /b/ and /p/ distributions of phonetic cues like voice onset time (VOT). We refer to this distributional learning as *phonetic adaptation*.

Like all perceptual category learning, phonetic adaptation can be *supervised* or *unsupervised*. In supervised learning, each observed VOT value is labeled with information that tells the listener whether the talker intended to produce /b/ or /p/. Such labeling information might come from, for instance, the surrounding word (*bash* vs. *pash*), or from visual cues to articulation. In unsupervised learning, however, no such labeling information is available. This is the case during language *acquisition* (e.g., Vallabha, McClelland, Pons, Werker, & Amano, 2007) but it can also occur in adult language adaptation when a VOT value occurs in a novel word, or a word that could have either /b/ or /p/, like *beach/peach*. In general unsupervised learning is harder: in addition to figuring out the *distribution* of VOTs for each category from limited observations, listeners also have to figure out how each of those observations should be categorized. Each of these depends on the other: how to categorize VOTs depends on the distributions for each category, while the distributions for each category depend on which VOTs are thought to belong to that category.

Both supervised and unsupervised phonetic adaptation have been observed in experiments. The earliest findings of phonetic adaptation were from supervised paradigms. For instance, after repeatedly hearing an ambiguous /f/-/s/ sound spliced into words that can only end in /f/ (e.g., *sheriff*), listeners classified more items on an /f/-/s/ continuum as /f/, and vice-versa when the ambiguous /f/-/s/ was spliced into /s/-final words (e.g., Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2005).

A small number of recent studies have demonstrated that phonetic adaptation can occur in an *unsupervised* context as well. Both Clayards, Tanenhaus, Aslin, and Jacobs (2008) and Munson (2011) had listeners listen to /b/-/p/ minimal pair words (e.g., *beach/peach*) with different VOTs, and click on a picture to indicate the word they heard. Across trials, the VOTs were drawn from a bimodal distribution with a low and a high VOT cluster. Listeners learned these distributions, as reflected in how they classified the VOT continuum, both the location and slope of their category boundary.

Such unsupervised adaptation requires that listeners combine the cue distributions they actually observe with their prior knowledge about what distributions are typical across talkers (Kleinschmidt & Jaeger, 2015). If a listener hears words with VOTs that cluster around 0 ms and 40 ms, they can infer that the mean VOT for /b/ is 0 ms and for /p/ is 40 ms, and that their classification should switch from /b/ to /p/ around 20 ms. In the absence of labels, each cue value is in principle ambiguous, and listeners need to observe enough different cue values to infer the underlying clusters.

In actual experience, however, phonetic adaptation is rarely purely unsupervised or supervised, with a mix or labeled and unlabeled observations. This raises the question: do listeners take advantage of extra information provided by labeled observations in phonetic adaptation? Work on domain-general *semi-supervised* category learning suggests that learners can leverage labeled trials to make learning from unlabeled trials even more effective (Gibson, Rogers, & Zhu, 2013). Existing phonetic adaptation paradigms do not directly answer this question, being purely supervised or purely unsupervised. Moreover, it's possible that what appears to be supervised learning in phonetic adaptation actually reflects a combination of cue-combination and *unsupervised* learning (Kleinschmidt & Jaeger, 2011, 2015). In this paper, we investigate the effect of adding some labeled trials to an otherwise unsupervised phonetic adaptation paradigm. This allows us to compare unsupervised and semi-supervised adaptation in the same paradigm, and thus directly assess the role that labeling information might play in phonetic adaptation.
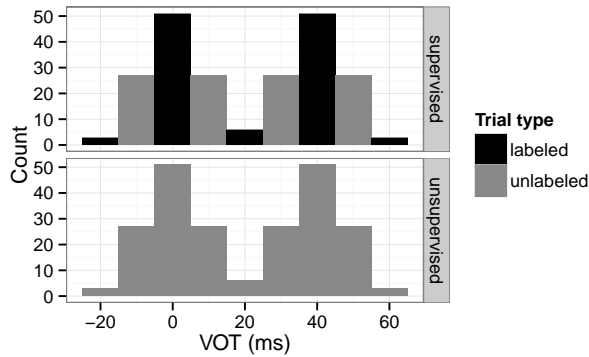
Figure 1: Stimuli distributions for unshifted condition in Experiment 1. The implied category boundary is at 20ms

Experiment 1

Methods

Subjects   We recruited 124 subjects via Amazon's Mechanical Turk, who were paid $2.00 for participation, which took about 20 minutes. We excluded subjects whose accuracy at 0 ms and 70 ms VOT—as extrapolated via a logistic GLM—was less than 80% correct. 10 subjects were excluded for this reason, leaving 114 for analysis.

Stimuli   Following Clayards et al. (2008), subjects heard spoken words, all members of /b/-/p/ minimal pairs (beach/peach, bees/peas, and beak/peak) synthesized with VOTs ranging from −20 ms to 90 ms. The actual VOT values that subjects heard were drawn from a bimodal distribution. The baseline, unshifted distribution (Figure 1) had a mean of 0 ms for /b/ and 40 ms for /p/ with an implied /b/-/p/ boundary at 20 ms. Subjects heard either this unshifted distribution, or a version that was shifted up by 10 ms VOT, with an implied category boundary at 30 ms VOT.

Procedure   On each trial, two pictures (target + distractor) were shown, and subjects were instructed to click on the picture that matched a spoken target word (e.g., *beach*). There were two kinds of trials. On *unlabeled* trials, the distractor picture was the minimal pair neighbor of the target word (e.g., a peach, Figure 2a), meaning that listeners had no additional information besides the VOT about whether the word started with a /b/ or a /p/. On *labeled* trials, the onset of the distractor picture's name was a minimal pair neighbor of the target word, but the rest was unrelated (e.g., bees, Figure 2b). This meant that the end of the word served as a label for the initial segment, and hence labeled the VOT value as either /b/ or /p/.

Subjects were randomly assigned to one of two conditions. In the *unsupervised* condition, all trials were unlabeled. In the *supervised* condition half were labeled and half unlabeled. In the supervised condition, each possible VOT was either always labeled, or always unlabeled (Figure 1). Specifically, the modal VOTs for /b/ and /p/ (0 ms and 40 ms in the unshifted condition) were always labeled, the stimulus at
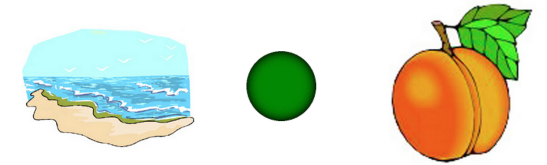
(a) Unlabeled trial with minimal pair distractor "peach".
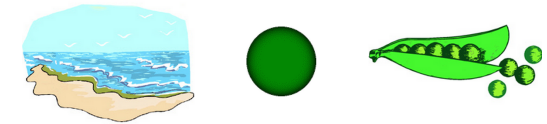
(b) Labeled trial with non-minimal pair distractor "peas".

Figure 2: Example trial displays for the target word "beach"

±10 ms VOT from the modal values (−10 ms, 10 ms, 30 ms, and 50 ms in the unshifted condition) were always unlabeled, and other stimuli were always labeled (−20 ms, 20 ms, and 60 ms).

Results

People used the labels for classification   On labeled trials in the supervised condition, listeners responded consistently with the label 98% of the time. This means that the response options available did, as we intend, effectively label the percept.

Learning was good overall   Figure 3 (top) shows the aggregate classification functions (averaged over subjects) for each third of the experiment. To evaluate how well listeners learned the distributions of VOTs they were exposed to, we analyzed the classification responses on unlabeled trials[1] using a mixed-effects logistic regression model. This model included fixed effects for stimulus VOT, supervised vs. unsupervised condition, distribution shift condition (0 ms or 10 ms), trial, and all interactions thereof. We used the maximal random effects structure for this design, with by-subject random intercepts and slopes for all the within-subject variables (trial, VOT, and their interaction). Table 1 shows the fixed effect coefficient estimates for this model and describes the details of how each variable was coded.

Figure 3 (bottom) shows the predictions of these fixed effects (i.e., the fitted classification functions) for each condition at the midpoint of each third of the experiment. We evaluated learning as the location of the /b/-/p/ category boundary, or where the fitted classification functions crossed the 50% /p/-response line.

Listeners learned well overall, and their classifications reflected the implied category boundaries of 20 ms and 30 ms within 2 ms.

---

[1]In the unsupervised condition, we only analyzed trials that would also have been unlabeled in the supervised condition.
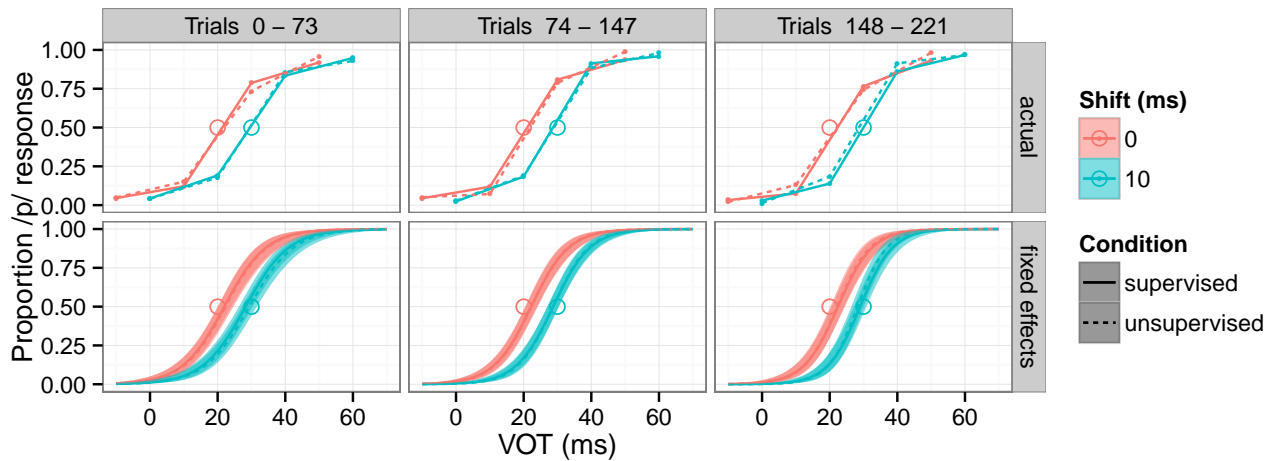
Figure 3: In Experiment 1, listeners' classification of unlabeled trials (lines) closely matches the implied category boundaries (open circles) for the unshifted (red) and 10 ms shifted (blue) distributions, but there is no difference between supervised and unsupervised learning (solid vs. dashed lines). Learning appears as the differences between 0 ms and 10 ms shifts (red vs. blue) and increasingly steep category boundaries (left to right). Top lines are raw average responses, and bottom lines are fitted logistic classification functions and 95% CIs on fixed effects (see Table 1).
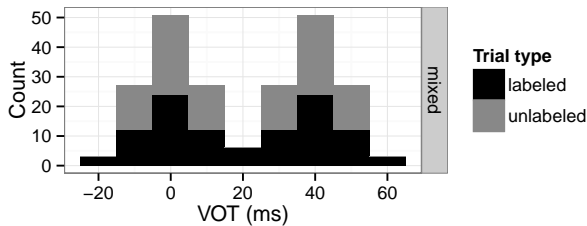

Figure 4: Stimuli distributions in Experiment 2, unshifted condition.

Supervision had no effect on learning Because labels reduce the difficulty of the distributional learning problem, we expected that learning would be faster or better overall in the supervised condition. Contrary to these expectations, learning in the supervised condition was neither faster, nor more complete, than in the unsupervised condition: the estimated category boundaries differ by less than 1 ms VOT between conditions.

Experiment 2

One of the shortcomings of the design of Experiment 1 is that in the supervised condition, listeners never heard exactly the same stimulus with and without a label. This means that the apparent inability or unwillingness of listeners to use the labels for learning might reflect stimulus-specific learning, as might be predicted by an episodic model of speech perception (Goldinger, 1998; Johnson, 1997). The sparse distribution of *unlabeled* trials may also reduce the statistical power by reducing the resolution with which the classification boundary can be estimated. Experiment 2 varies the design slightly

to determine whether labels affect adaptation when the same stimuli occur as labeled and unlabeled, and when unlabeled test trials occur over a broader range of VOTs.

Methods

The design was identical to that of Experiment 1, except for the following modifications. First, we modified the supervised condition, spreading out labeled and unlabeled trials more evenly (see Figure 4). Across trials, many VOT values occurred as both labeled and unlabeled trials, unlike in the supervised condition of Experiment 1 where each VOT value only occurred as labeled, or only occurred as unlabeled. Second, we only ran this modified supervised condition, and compared it to the unsupervised condition of Experiment 1.

Subjects We recruited 62 subjects via Amazon's Mechanical Turk, who were paid $2.00 for participation, which took about 20 minutes to complete. 2 subjects were excluded for failing to reliably classify the continuum, and 2 were excluded from analysis because they had already participated in Experiment 1, leaving 58 subjects for analysis.

Results

As in Experiment 1, on labeled trials listeners used the labels to guide their responses, responding consistently with the label 98% of the time.

We analyzed learning in the same way as Experiment 1, using the unsupervised condition from Experiment 1 as a baseline. Unlike in the analysis of Experiment 1, we considered all trials from the unsupervised condition, because the labeled trials in the supervised condition of Experiment 2 covered the entire continuum. Figure 5 shows the raw data (top) and the fitted classification functions (bottom) and Table 1 shows the
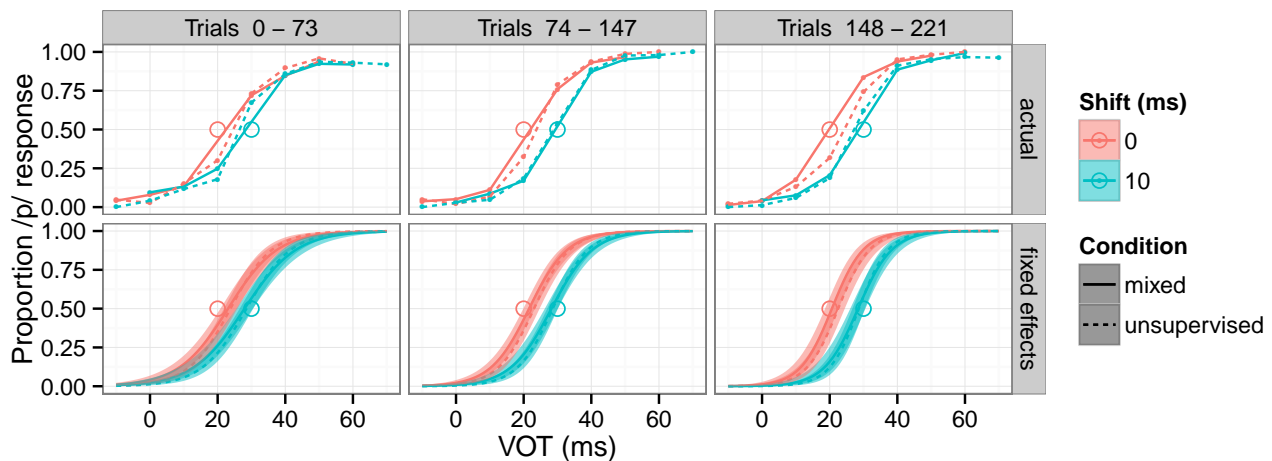
Figure 5: In Experiment 2, for both distributions listeners' classification (lines) closely matches the category boundary implied by the distributions (open circles), just as in Experiment 1 (compare with Figure 3). Labels still made no difference (solid vs. dashed lines), even though labeled trials were distributed more evenly over the VOT continuum than in Experiment 1.

fixed effects estimates.

As in Experiment 1, listeners learned quickly and their category boundaries were very close to those implied by the distributions of VOTs they heard. Again, however, learning in the supervised condition (of Experiment 2) was neither faster nor more complete than in the unsupervised condition (of Experiment 1): the category boundaries for supervised and unsupervised were within 2 ms of each other.

Discussion

Even when the same stimuli occur with and without labels, the availability of labels appears to make little difference in adapting to a novel talker's /b/ and /p/ categories. This suggests that the failure to find effects of supervision in Experiment 1 was not due to the fact that labeled and unlabeled stimuli were acoustically different.

General Discussion

In two experiments we directly compared phonetic adaptation with and without supervision. The presence of information that labels an acoustic stimulus as a /b/ makes the task of learning the distribution of acoustic cues for the /b/ category easier, at least in principle. Normative theories that treat phonetic adaptation as a kind of distributional learning thus predict that, in general, the availability of labels should make adaptation faster, more complete, or both (Kleinschmidt & Jaeger, 2015).

Contrary to this prediction, we did not find any effect of supervision on the distributional learning of cue-category mappings in adults. At first glance this contradicts the results of other studies on supervised phonetic adaptation, which suggest that people *do* use labeling information to facilitate learning. For instance, Norris et al. (2003) found adaptation when listeners heard an ambiguous /s-ʃ/ spliced into words that

consistently labeled it as either /s/ or /ʃ/. However, when Norris et al. (2003) spliced the same sound in *novel* words that provided no labeling information, listeners did not adapt, suggesting that labeling is crucial for phonetic adaptation. How can we reconcile these apparently contradictory results? We briefly discuss four possibilities here: that the *kind* of label matters, that learning was too easy, that self-supervision overwhelms any outside labels in this task, and that our labels were not sufficiently informative.

What kind of label?

One possibility is that the *kind* of label matters. In previous studies on phonetic adaptation where labels are provided, the labels come either from a visual component of the stimulus (e.g., a video of a natural production of /aba/, dubbed over audio of an ambiguous /aba/-/ada/, Bertelson, Vroomen, & de Gelder, 2003) or from the lexical context (e.g., an ambiguous /s/-/ʃ/ spliced into the word *dino_aur*, Norris et al., 2003; Kraljic & Samuel, 2005). In both cases, the label is an intrinsic part of the (audio-visual) speech signal itself. In our design, the label comes from the pragmatic context, the available response choices. It is possible that listeners can use this sort of pragmatic information to guide their responses, but that it is nevertheless not available to whatever systems are responsible for perceptual learning.

A related possibility is that labels that are intrinsic to the signal affect distributional learning in a purely bottom-up way. That is, disambiguating visual information (a natural video of /aba/) might function not at the level of identifying the *category* that the talker intended to produce, but by changing the *cue* that is perceived. Indeed, there is abundant evidence that cues are combined in this way within and across modalities, in speech perception (Bejjanki, Clayards, Knill, & Aslin, 2011; Toscano & McMurray, 2010) and in perception

more generally (cf. Ernst & Bülthoff, 2004). If adaptation is driven by distributional learning of the integrated multimodal percept, rather than the component cues, then what appears to be sensitivity to category labels in previous adaptation studies may instead by bottom-up distributional learning of not-fully-ambiguous multimodal cues (Kleinschmidt & Jaeger, 2011).

A learning ceiling effect?

Listeners adapted very well to both the unshifted and 10 ms-shifted distributions, with their classifications matching the implied category boundaries even in the first third of the experiment. This suggests that learning these distributions may have been too easy for the labels to make any difference.[2] It remains a question for future work to see whether a more sensitive paradigm can find an effect of labels by, for instance, using a smaller number of exposure trials to induce adaptation coupled with with a separate pre- and post-test to assess adaptation.

Self-supervision

Unlike most studies on domain-general semi-supervised learning, listeners in our studies have a great deal of prior experience with the categories we are teaching them, at least as they are produced by other talkers. This makes even our unsupervised condition partially supervised: listeners' prior experience provides a *self*-supervision signal, or, in Bayesian terms, a prior (Kleinschmidt & Jaeger, 2015). It thus could be the case that this prior is sufficiently informative to make any additional information provided by the labels themselves redundant.

A related, if more extreme, possibility is that listeners might decide the first time they hear, for instance, a VOT of 10 ms that it is a /b/, and never change that belief. However, the fact that the category boundaries grow *steeper* with more exposure suggests that this is not correct: if listeners committed to a categorization of each individual stimulus early, then their categorization functions should be sharp and constant throughout the experiment.

How informative is each label?

Previous phonetic adaptation studies that used labels applied those labels to trials that were acoustically maximally ambiguous (e.g., Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris et al., 2003). This makes each label maximally informative without causing a cue conflict between the label and how listeners would have classified the cue without a label. In our design, labels occurred on many different cue values, many of which listeners would already have classified consistently with the label a priori. Thus, on average, each label in our design provides substantially less information for the listener than in previous designs. This may explain the failure to find any effect of labels on adaptation: listeners simply

did not gain enough extra information about the underlying distributions from the labels we provided them for it to make a difference in their learning behavior. This possibility seems the most likely explanation of our results, and calls for further work using the same *kind* of labels, but with shorter exposure where the labels are more informative along the lines of earlier supervised adaptation studies (e.g., Norris et al., 2003).

Conclusion

In two studies, we found that phonetic adaptation was insensitive to label information, even thought those labels changed listeners' classifications. Normative theories that see phonetic adaptation as a sort of statistical inference predict that listeners should use all information available to them in order to more effectively adapt to novel talkers (Kleinschmidt & Jaeger, 2015). While our results appear to violate that prediction, there are some important caveats. Most importantly, the labels we used may not have provided enough additional information about the underlying distributions, and for the purposes of learning the category distributions may have been redundant with the statistics of the cues themselves. This suggests a more nuanced understanding of the predictions of normative models of adaptation. The combination of prior experience with other talkers and sufficient observations from a category might mean that, in many everyday situations, the availability of labels does not contribute enough extra information to change listeners' behavior. Further modeling and behavioral work is required to investigate the tradeoff between prior experience, number of observations, and informativity of labels in adaptation. Regardless, it is still important to note that the same labels may be informative about how to classify but relatively uninformative about the overall distribution.

Acknowledgments

References

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE*, *6*(5), e19812. doi: 10.1371/journal.pone.0019812

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi: 10.1016/j.cognition.2008.04.004

---

[2]We also investigated larger shifts of 20 ms and 30 ms, for which adaptation was incomplete. Nevertheless, labels made no difference and so for the sake of brevity we do not report the detailed results here.

	Experiment 1	Experiment 2
(Intercept)	−0.08 (0.08)	−0.09 (0.08)
VOT	1.83*** (0.08)	1.83*** (0.07)
Shift	0.65*** (0.16)	0.75*** (0.16)
Supervised	0.001 (0.16)	0.16 (0.15)
Trial	−0.02 (0.16)	0.17 (0.18)
VOT : Shift	0.05 (0.14)	−0.03 (0.14)
VOT : Supervised	−0.06 (0.14)	−0.19 (0.13)
Shift : Supervised	−0.02 (0.32)	−0.27 (0.30)
VOT : Trial	0.76*** (0.15)	1.01*** (0.14)
Shift : Trial	0.35 (0.27)	−0.11 (0.33)
Supervised : Trial	−0.34 (0.27)	0.13 (0.32)
VOT : Shift : Supervised	−0.04 (0.29)	−0.13 (0.26)
VOT : Shift : Trial	0.36 (0.25)	0.18 (0.25)
VOT : Supervised : Trial	−0.22 (0.25)	0.12 (0.25)
Shift : Supervised : Trial	−0.40 (0.54)	−0.38 (0.64)
VOT : Shift : Supervised : Trial	−0.25 (0.50)	−0.49 (0.49)
Observations	12,312	18,378

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 1: Fixed effect coefficients (and standard errors) for mixed effects regression models of data from Experiments 1 and 2. All categorical predictors were sum-coded (with range normalized to 1). To minimize collinearity between distribution shift and stimulus VOT, stimulus VOT was re-coded relative to the implied category boundary. This means that the VOT predictor was uncorrelated with the distribution shift predictor. To improve convergence, the VOT and boundary shift predictors were coded as continuum steps (divided by 10) to put them on roughly the same scale as the other predictors. Finally, trial number was centered and scaled to a range of 1 (very first trial = −0.5, very last trial = 0.5).

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–9. doi: 10.1016/j.tics.2004.02.002

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*(1), 132–72. doi: 10.1111/tops.12010

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–79.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Kleinschmidt, D. F., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd acl workshop on cognitive modeling and computational linguistics.* Stroudsburg, PA: Association for Computational Linguistics. Talk.

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). doi: 10.1037/a0038695

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–78. doi: 10.1016/j.cogpsych.2005.05.001

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing*. Unpublished doctoral dissertation, University of Iowa.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi: 10.1016/S0010-0285(03)00006-9

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13273–8. doi: 10.1073/pnas.0705369104