frontiers

# Using rational models to understand experiments on accent adaption

**Maryann Tan** [1,2,†,*], **Xin Xie** [2,†] **and T. Florian Jaeger** [2,3]

[1] *Centre for Research in Bilingualism, Dept. of Swedish & Multilingualism, University of Stockholm, Stockholm, Sweden*
[2] *Brain & Cognitive Sciences, University of Rochester, Rochester, NY, USA*
[3] *Computer Science, University of Rochester, Rochester, NY, USA*
[†] *These authors share first authorship*

Correspondence*:
Maryann Tan
Centre for Research in Bilingualism
Stockholm University
SE-106 91 Stockholm, Sweden
maryann.tan@biling.su.se

## ABSTRACT

Exposure to unfamiliar non-native speech tends to improve comprehension. One hypothesis holds that listeners adapt to non-native-accented speech through distributional learning—by inferring the statistics of the talker's phonetic cues. Models based on this hypothesis provide a good fit to incremental changes after exposure to atypical *native* speech. These models have, however, not previously been applied to non-native accents, which typically differ from native speech in many dimensions. Motivated by a seeming failure to replicate a well-replicated finding from accent adaptation, we use ideal observers to test whether our results can be understood solely based on the statistics of the relevant cue distributions in the native- and non-native-accented speech. The simple computational model we use for this purpose can be used predictively by other researchers working on similar questions. All code and data are shared.

**Keywords: L2 speech, non-native speech, foreign accent, adaptation, distributional learning, ideal observer**

## 1 INTRODUCTION

Understanding strongly non-native-accented speech can be challenging: native listeners unfamiliar with a non-native accent tend to process it more slowly and with decreased accuracy (Munro and Derwing, 1995; Witteman et al., 2013). There is now ample evidence that this initial processing disadvantage can decrease with exposure to the accented talker (e.g. Bradlow and Bent, 2008; Adank et al., 2009; Weil, 2001), with some improvements emerging within mere minutes of exposure (Clarke and Garrett, 2004; Xie et al., 2018b). What has remained less well understood are the mechanisms underlying these changes in speed and accuracy of processing.

Two broad classes of (mutually compatible) hypotheses have emerged. One holds that changes in native listeners' processing of non-native-accented speech arise from a general relaxation of decision criteria for phonological categorization (e.g., "general expansion", Schmale et al., 2012). The other hypothesis holds that listeners learn talker- or even accent-specific characteristics, including information about specific segmental features and super-segmental properties of the accented speech (e.g., Bradlow and Bent, 2008;

Sidaras et al., 2009). This latter hypothesis has received further elaboration: that adaptation to non-native accents is at least in part achieved through distributional learning (Wade et al., 2007; Kartushina et al., 2016; Idemaru and Holt, 2011; Schertz et al., 2015) of the type assumed in exemplar (Pierrehumbert, 2001) or Bayesian theories of speech perception (Kleinschmidt and Jaeger, 2015).

Distributional learning models have been found to provide a good qualitative and quantitative explanation of certain adaptive changes listeners exhibit in response to shifted or otherwise atypical pronunciations by native talkers (Bejjanki et al., 2011; Clayards et al., 2008; Kleinschmidt and Jaeger, 2015, 2016; Theodore and Monto, 2019). This includes changes in categorization boundaries observed in perceptual recalibration (e.g. Eisner and McQueen, 2005; Drouin et al., 2016; Kraljic and Samuel, 2006; Norris et al., 2003) or unsupervised learning paradigms (Clayards et al., 2008; Nixon et al., 2016). However, tests of distributional learning models have almost exclusively been limited to comparatively small deviations from the expected means or variances of two phonological categories along a single phonetic dimension (for examples with two phonetic dimensions, see Hitczenko and Feldman, 2016; Xie et al., 2021). Whether distributional learning can explain adaptation to the types of more complex deviations from expected pronunciations that are observed in unfamiliar non-native accents is an open question. Specifically, non-native accents differ from the expected native pronunciation along many acoustic and linguistic dimensions, including both supra-segmental and segmental differences. Non-native speech might, for example, realize segmental or supra-segmental categories with means that are shifted relative to native means (Best, 1995; Flege, 1995) and with expanded or reduced variance (Smith et al., 2019; Vaughn et al., 2019; Xie and Jaeger, 2020), including deviation in terms of the relative reliance on different cues to signal the same phonological contrast (Flege et al., 1992; Xie et al., 2017). In short, adaptation to a talker with an unfamiliar non-native accent constitutes a more complex problem than adjustments in response to more limited differences between native talkers, and it is possible that these challenges require a different set of mechanisms (for related discussion see Goslin et al., 2012; Porretta et al., 2017).

We take a hugely simplified step towards addressing this question. Our approach is post-hoc and confirmatory (although future work might employ the same approach *pre*dictively prior to data collection). We ask whether a simple model of speech perception (an ideal observer, Clayards et al., 2008; Kleinschmidt and Jaeger, 2015; Norris and McQueen, 2008) can be employed to make informative predictions as to whether exposure to a specific set of non-native-accented speech stimuli is expected to result in detectable adaptation (see also Hitczenko and Feldman, 2016). To demonstrate the potential value of such an approach, we ask whether an ideal observer sheds light on what appeared to be, at first blush, a failure to replicate previous findings from accent adaptation (Eisner et al., 2013; Xie et al., 2017), despite very similar design and procedure.

We emphasize that our goal here is not to convincingly argue that distributional learning is the best explanation for the data at hand. Rather, we aim to demonstrate *how* one can use a simple normative model of speech perception to derive predictions for the perception of, and adaptation to, non-native-accented speech. By comparing the responses of human listeners to the predictions of this computational model, researchers can achieve a clearer sense of which results (null or not) should be treated as surprising (see also Massaro and Friedman, 1990, on the value of normative models for speech perception). While models of speech perception suitable for this purpose now exist (Clayards et al., 2008; Kleinschmidt and Jaeger, 2015), they are still rarely employed in the interpretation of experimental results (but see e.g., Lancia and Winter, 2013; Kleinschmidt et al., 2015; Hitczenko and Feldman, 2016; Theodore and Monto, 2019; Xie et al., 2021). The present report aims to demonstrate how even the post-hoc application of computational models to experimental data can aid interpretation. It also holds the potential to reduce the "file drawer" problem

(Rosenthal, 1979)—the bias to not publish null results—as well as to pre-empt the 'over-interpretation' of null results. As we illustrate below, not every null result is a Type II error; null results can be precisely what a model predicts given the specific stimuli of an experiment. We thus hope this report can serve as a helpful guide, encouraging experimenters to interpret results with regard to more fully specified models. To this end, this report is accompanied by detailed supplementary information (SI) written as executable, richly documented, R markdown (Allaire et al., 2021) and compiled into an interactive HTML. These SI, along with all data, are shared via the Open Science Framework (https://osf.io/72fkx/). The main text aims to provide a high-level overview of the approach and results.

## 2 THE 'PUZZLE'

The two perception experiments we aim to understand share the same exposure-test design and procedure (Figure 1), but differ in the L1-L2 pair investigated. Both experiments investigate adaptation to non-native-accented speech with regard to the same phonological category—syllable-final voicing of /d/—present in the non-native language, but absent in the native language. The pronunciation of syllable-final /d/ is known to be affected in both non-native accents investigated.

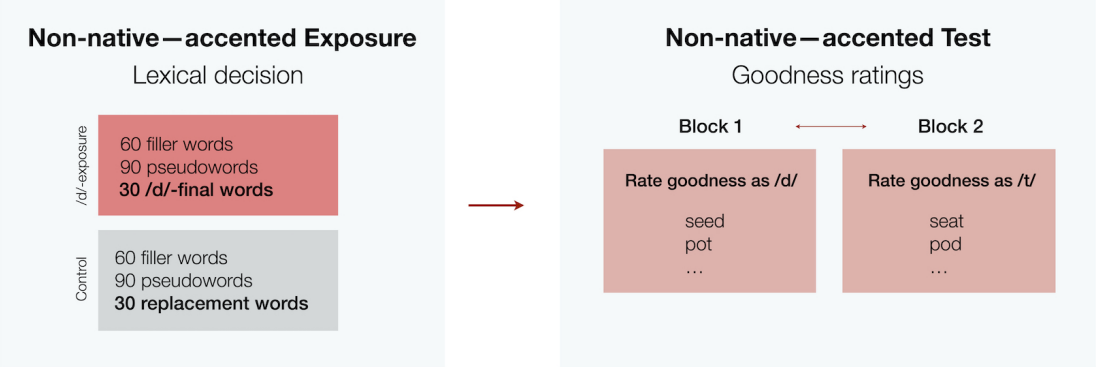


**Figure 1.** Design of English and Swedish experiment analyzed here. The order of /d/- and /t/-goodness test blocks was counter-balanced across participants.

The first experiment exposed native speakers of American English to Mandarin-accented English speech (Xie et al., 2017). Unlike English, Mandarin does not have stops in syllable-final position. As would be expected from theories of L2 learning (e.g. Flege, 1995), the realization of final stop-voicing differs between native English and Mandarin-accented English (Flege et al., 1992; Xie and Jaeger, 2020). This was also confirmed specifically for the non-native-accented speech materials used in the experiment (Xie et al., 2017).

Exposure was manipulated between participants. Both groups heard 90 words and 90 pseudowords while conducting a lexical decision task. For the */d/-exposure* group, this included 30 words containing a syllable-final /d/ (e.g., *lemonade*). These exposure words were chosen to not have minimal pair neighbors with syllable-final /t/, allowing lexical guidance on the non-native talker's /d/ productions. Participants in the *control group* heard no words with syllable-final /d/ (for details about the materials, see SI). Neither groups heard syllable-final /t/ productions during exposure.

During test, participants in both groups heard the same minimal pair words with syllable-final /d/ or /t/ (e.g., a recording of *seed* or *seat*). Participants had to rate how "good" the word sounded as an instance of

/d/ (one block) or /t/ (another block, with the order of blocks counter-balanced across participants). Words within the same minimal pair did not appear in the same block (see Figure 1).

Goodness ratings have been used to analyze listeners' representations of the internal structure of phonological categories (e.g. Samuel, 1982; Volaitis and Miller, 1992; Allen and Miller, 2001), including after exposure to shifted native categories in perceptual recalibration (e.g., Drouin et al., 2016). Xie et al. (2017) found that /d/-exposure lead to improved goodness ratings for the non-native-accented /d/- and /t/-final words during test, compared to the control group. We refer to this as the English data. Xie and colleagues replicated the effect of /d/-exposure in three additional experiments using the same recordings and similar exposure-test paradigms but different tasks and participants (Xie et al., 2017; Xie and Myers, 2017; Xie et al., 2018a). The same effect has also been found in experiments with similar designs on syllable-final /d/ in Dutch-accented English, which tends to devoice final stops (Eisner et al., 2013).

In a recent experiment however, we failed to find the effect of /d/-exposure for another L1-L2 pair, Flemish-accented Swedish. Unlike Swedish, Flemish (a dialect of Dutch) devoices voiced stops in syllable-final position (Booij, 1999; Verhoeven, 2005). This type of phonological rule is well-documented to transfer from a talker's first language to their second language and was confirmed in the L2-accented speech materials used in the Swedish experiment (Tan et al., 2019). Like with Dutch- and Mandarin-accented English, we thus expected exposure to Flemish-accented Swedish syllable-final /d/ to affect ratings during test. Both the English and Swedish experiments used lexically-guided exposure with the same task. Both experiments manipulated exposure to the non-native-accented sound (syllable-final /d/) in the same two between-participant conditions, including the same amount of exposure. Both experiments used /d/ and /t/ goodness ratings of /d/-/t/-final minimal pair words during test. Unlike Xie et al. (2017), however, the Swedish data did *not* yield an effect of /d/-exposure on ratings during test. In fact, the effect of exposure went numerically in the opposite direction in the Swedish data.

Figure 2 (top) shows the rating results from both experiments. Linear mixed-effects regression presented in the SI (§3.2.2) confirmed that the effects of exposure differed significantly between the two experiments (coefficient-based $t$-test, $p < .002$): whereas /d/-exposure resulted in significant facilitation for English ($\hat{\beta} = .04$, $p < .0001$), it did not for Swedish—in fact, trending in the opposite direction ($\hat{\beta} = -.03$, $p > .1$).

At first blush, the Swedish data seem to constitute a failure to replicate the English experiment. In particular, since the effect found in the English data has been replicated a number of times, it would be tempting to consider the Swedish result a Type II error (rather than the English result a Type I error). Further, adding to this interpretation, the Swedish experiment collected substantially less data: while the English data consists of 120 ratings each from 48 participants, the Swedish data consist of 60 ratings each from 23 participants—about a fourth of the English data. This would seem to suggest lack of statistical power as a straightforward explanation for the null effect in the Swedish experiment. However, even when the English data was down-sampled to the size and structure of the Swedish data, the difference between the two data sets remained significant 57.6% of the time (out of 1000 hierarchical bootstrap samples, SI, §3.2.5). For English, the simple effect of /d/-exposure went in the predicted direction 89.6% of the time, reaching significance in 44.4% of all bootstrap samples (vs. 0.6% significant effects in the opposite direction). For Swedish, the simple effect went in the predicted direction 29.2% of the time, and was significant in 7.6% of all samples (vs. 40.2% significant effects in the opposite direction).

Overall, this suggests that power differences alone are unlikely to fully explain the difference between the English and Swedish results. Indeed, the same hierarchical bootstrap analyses found that the Swedish

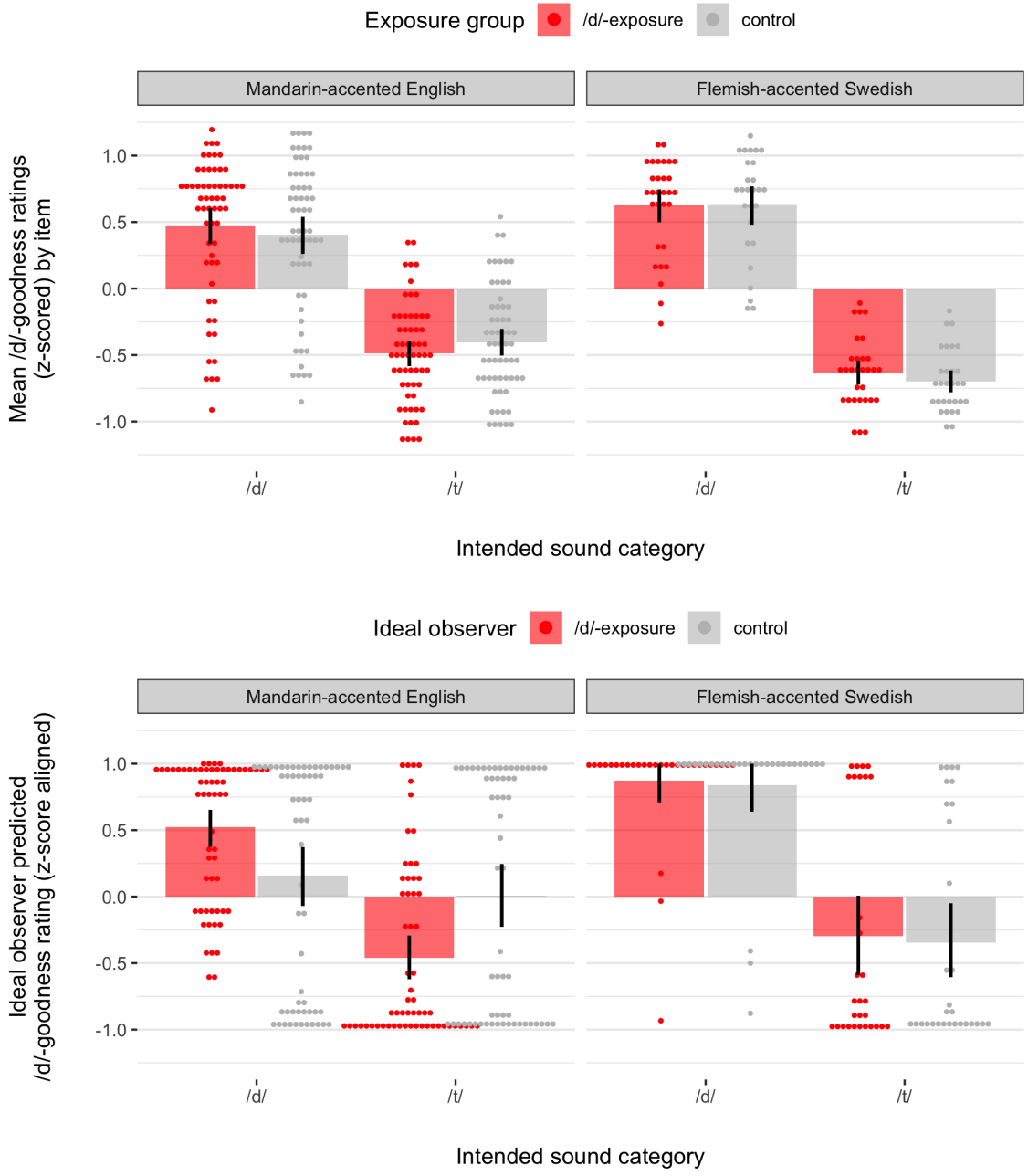


**Figure 2.** Top: Results of behavioral experiment on native listeners' perception of syllable-final /d/ and /t/ in Mandarin-accented English (left) and Flemish-accented Swedish (right). Points show by-item means of z-scored /d/-goodness ratings (standardized within each participant) for non-native productions of syllable-final /d/ and /t/ during test, depending on the whether participants received exposure to the relevant non-native realization of syllable-final /d/ (/d/-exposure) or not (control). Bars show means and 95% bootstrapped confidence intervals of the by-item means. Bottom: Ideal observer-predicted /d/-goodness ratings described in Section 3.2.

results are very unlikely to result if the English experiment is taken as the 'ground truth': only 12 out of 1000 (1.2%) random resamples of the English experiment resulted in $t$-values as small or smaller than the one observed in the Swedish experiment.

What then caused the difference in results? And do the Swedish data really constitute a Type II error? The SI (§2) discusses a comprehensive list of differences in methodology between the experiments. This comparison revealed that the recordings for two experiments had been obtained in different ways. The Flemish-accented Swedish materials were elicited by first playing a native-accented recording of the word, whereas the Mandarin-accented English materials were elicited without such assistance (SI, §2.2). This raised the possibility that the Flemish-accented Swedish recordings deviated less from native Swedish than the Mandarin-accented English recordings deviated from native English, which would reduce the perceptual benefit of /d/-exposure.

An initial comparison of the non-native-accented /d,t/ productions during test to productions of the same test words by a Swedish native speaker (not included in the experiment, but recorded using a similar procedure) lends credence to this hypothesis. Figure 3 shows native- and non-native-accented syllable-final /d,t/ productions of all test items for both English and Swedish. Native productions were obtained from one or more gender-matched speakers similar in age to the non-native speakers employed in the experiments (for details, see SI, §2.2.1). We annotated native- and non-native-accented production for three cues known cross-linguistically to signal syllable-final stop voicing: the duration of the preceding vowel, the duration of the closure interval, and the duration of the burst release (for details on the annotation procedure, see SI, §4.1). Syllable-final stop voicing in Mandarin-accented English is known to differ in the use of these three cues, compared to native-accented English (Xie and Jaeger, 2020), as also clearly visible in the left panels of Figure 3 (replicating Xie et al. 2017). At least superficially, the Flemish-accented recordings seem to deviate less strongly from the native Swedish productions (right panels) than the Mandarin-accented recordings deviate from native English productions (left panels).

In line with this initial impression, the Flemish-accented Swedish recordings were substantially easier to process for the Swedish participants compared to the Mandarin-accented English recordings for the English participants: lexical decision accuracy during exposure was substantially higher for the Swedish data (Swedish, d-exposure: 96%, control: 97%) than for the English data (/d/-exposure: 78%, control: 74%). This included accuracy on the critical exposure words with syllable-final /d/ (English, /d/-exposure: 78%, SD = 9%; Swedish, /d/-exposure: 94%, SD = 6%; for further detail, see SI, §3.1).[1]

We thus decided to estimate the predicted consequences for the benefit of /d/-exposure for each experiment given the specific distributional properties of (1) the non-native-accented /d/ in the /d/-exposure group in that experiment, (2) the 'typical' native-accented /d/ and /t/ in that language, and (3) the non-native-accented /d,t/-final minimal pair words during test. From this point on—having ruled out a number of alternative explanations for the seemingly diverging results—our approach is confirmatory: our goal is not to rule out alternative mechanisms for accent adaptation but rather to explore how a simple but fully specified computational model of distributional learning can aid data interpretation. This, we hope, may be informative for researchers who find themselves in a situation similar to the one described here: trying to understand (or even predict) the results of an experiment—specifically, the expected results based on the distributional properties of the speech stimuli employed in the experiment.

[1] The difference in exposure accuracy could also be explained if the Swedish participants were more familiar with accents that involve syllable-final devoicing than the American participants. For example, exposure to German-accented Swedish is common in Stockholm (as our Swedish colleagues were eager to point out). Post-experiment surveys found that none of the Swedish participants was able to guess the L1 of the non-native accent, and only one (4.3%) of the participants guessed another L1 that leads to syllable-final devoicing (German). It is possible, however, that participants nevertheless had subconscious familiarity with syllable-final devoicing. This would explain the exposure results. It would not, however, explain the differences in the degree of accentedness in the *productions*, shown in Figure 3.

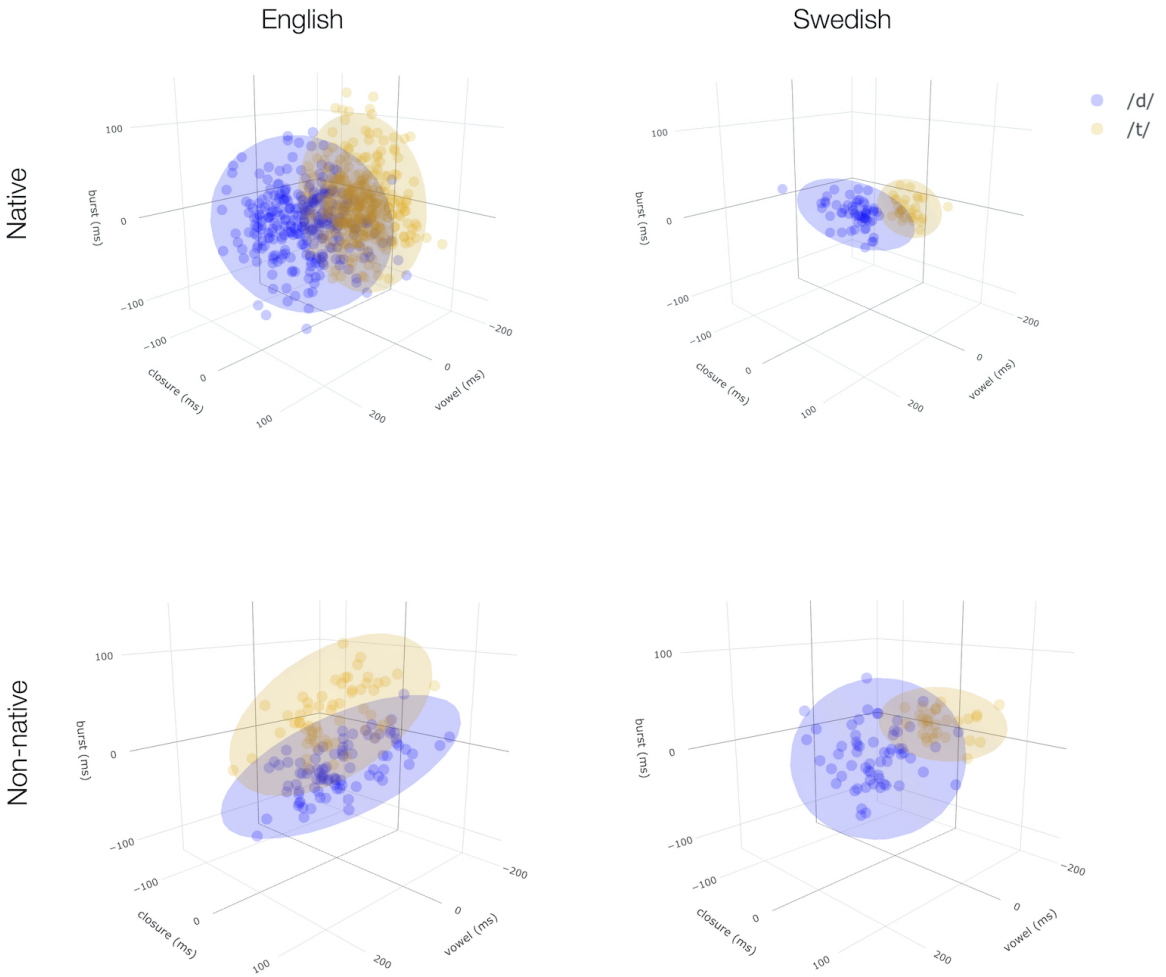


**Figure 3.** Comparison of native- (top) and non-native-accented (bottom) syllable-final /d,t/ for both the English (left) and the Swedish (right) data. Productions combine information from multiple databases and are corrected for phonotactic context effects (see SI, §4.3) and are shown in the 3D space defined by three important cues (duration of vowel, closure, burst) to syllable-final voicing. Ellipses contain 95% of the probability density under the assumption that categories form multivariate Gaussian distributions. To facilitate comparison, axis limits are held constant across panels. See SI (§4.3.2) for interactive visualization.

## 3 MODELING THE EFFECT OF EXPOSURE

We approach this question using ideal observers, specifically ideal categorizers, though we note that exemplar models (e.g., Shi et al., 2010) would make similar predictions for the present purpose. We use ideal observers because they provide an analytic framework to derive how an ideal/rational listener should respond to input given a certain set of assumptions (for early discussion of the value of this approach, see Massaro and Friedman, 1990). Like exemplar models, ideal observers emphasize the importance of categories' cue distributions. Specifically, the posterior probability of recognizing an input as category $c$ is a function of both the category's prior probability, $p(c)$, and the probability of observing the input under the hypothesis that the speaker intended to produce category $c$ (the "likelihood"), $p(cues|c)$. These two pieces of information are assumed to be integrated optimally, as described by Bayes' theorem:

$$p(c|cues) = \frac{p(cues|c) * p(c)}{\Sigma_i p(cues|c_i) * p(c_i)} \quad (1)$$

Of appeal is that categorization in this approach has zero computational degrees of freedom, and yet has been found to provide a good explanation for a variety of phenomena in speech perception and spoken word recognition (e.g., Bejjanki et al., 2011; Clayards et al., 2008; Feldman et al., 2009; Kronrod et al., 2016; Luce and Pisoni, 1998; Kleinschmidt and Jaeger, 2015; Norris and McQueen, 2008).

Here we use ideal observers as a methodological tool to estimate how an idealized participant who has adapted to the phonetic distributions in the input during exposure would respond to the test items. The model makes a number of simplifying assumptions—many of them known to be wrong, but none of them trivially explaining the predictions we derive. These assumptions are summarized in Table S1 in the supplementary materials. Here we emphasize only the assumptions that make the models ideal*ized* rather than *ideal* (for the same distinction, see also Qian et al., 2016): rather than model ideal incremental adaptation to the exposure stimuli (Kleinschmidt and Jaeger, 2015), we model listeners that (1) have *completely* adapted by the end of exposure, and (2) do not adapt further during the test phase or at least not much. While (2) is plausible (inputs during test are not lexically labeled since they are minimal pair words; and adaptation seems to proceed most quickly upon initial exposure to talkers, Kraljic and Samuel 2007), assumption (1) is likely wrong. Indeed, ideal adaptation should weight and integrate the observed input from a talker with prior expectations, so that only partial adaptation is expected after exposure to 30 critical words—partial in the sense that listeners' representations are not a replica of the statistics of the non-native speech, but rather somewhere between the native and non-native speech (Kleinschmidt and Jaeger, 2015).

### 3.1 Methods

We developed four ideal observer models, matching the four combinations of experimental conditions: 2 experiment (Swedish vs. English) X 2 exposure group (/d/-exposure vs. control). Our goal was to approximate the effects of exposure in these four conditions. All models encode listeners' beliefs about /d/ and /t/ as multivariate Gaussian distributions in the 3D space defined by vowel, closure, and burst duration.

To approximate the effect of /d/-exposure, we estimated the mean and covariance of the /d/ category from the 30 non-native-accented recordings of the syllable-final /d/ employed during the experiments' exposure phase. To approximate the effect of control exposure, we estimated the mean and covariance of the /d/ category from recordings of the same 30 exposure words by a gender- and age-matched native speaker. Since by design, neither /d/- nor control exposure contained similarly lexically-labeled instances of syllable-final /t/, we made the simplifying assumption that both idealized listeners would have native /t/ categories. This ignores that listeners might adapt their expectations about /t/ based on exposure to the talker's /d/ or other categories whose realization is correlated with that of the /t/ category (see, e.g., Chodroff and Wilson, 2017). The SI describes the databases (§4.1) and annotation procedure (§4.2) we employed to estimate the means and covariances of the native /t/ and non-native /t/ and /d/ categories.

While test words formed minimal pairs, holding phonotactic context constant across productions of /d/ and /t/, this was not the case between exposure and test productions. We thus use multiple linear regression to correct cue values for effects of segmental, supra-segmental and talker context (for details, including interactive plots illustrating the consequence of the correction procedure, see SI, §4.3). This approach closely follows the influential C-CuRE model of cue normalization (McMurray and Jongman, 2011), extending it to the contrast between native and non-native speech. C-CuRE has been found to provide a good fit against human categorization responses, including influences of coarticulation due to

phonotactic context (Apfelbaum and McMurray, 2015). All ideal observers were fitted to and evaluated on these context-corrected cue values (SI, §4.4).

Both the native and non-native ideal observers were then applied to the *non-native-accented* minimal pair words from the test phase of the experiments (SI, §4.5). For each test token, we calculated the ideal observer's posterior probability of /d/ (and /t/), using Bayes theorem. In order to relate the posterior probabilities of /d/ and /t/ to participants' goodness ratings, it is necessary to specify a linking hypothesis. Conveniently, human categorization responses for the same stimuli and the same exposure conditions as analyzed here are available from a separate experiment in Xie et al.. Paralleling Xie and colleagues' rating experiment, the categorization experiment found the predicted shift in the /d/-/t/ category boundary following /d/-exposure, compared to control exposure (Xie et al., 2017). This allowed us to investigate the relation between human goodness ratings and proportions of categorization responses, using generalized additive mixed models (GAMMs, Hastie, 2017). These analyses (presented in the SI§4.5.4) revealed a clearly linear relation between proportion /d/-responses in categorization and /d/-goodness ratings (and, vice versa, for /t/), at least for the type of stimuli analyzed here. For our analyses, we thus assume a simple identity link between the ideal observers' predicted posterior probability of a category and listeners' goodness ratings for that category. For visualizations (e.g., Figure 2, bottom), we facilitate comparison of ideal observers' prediction to human ratings by scaling the ideal observer-predicted posterior probabilities (range = 0 to 1) to have the same range as human rating responses across the combined English and Swedish data (range = -1 to 1). In those visualizations, we refer to the resulting predictions as posterior ratings. This scaling does not affect correlations between the ideal observers' predictions and human rating responses.

### 3.2 Results: Goodness ratings predicted by ideal observer

Figure 2 (bottom row) shows the results for the control and /d/-exposure ideal observers and both exposure conditions. Paralleling participants' goodness ratings for Mandarin-accented English in Figure 2, posterior ratings were improved under the non-native English model compared to the native English model. And, paralleling participants' goodness ratings for Flemish-accented Swedish, no such improvement of posterior ratings was observed under the non-native Swedish model compared to the native Swedish model.

The ideal observers thus predict effects of exposure condition on goodness ratings that *qualitatively* resemble the results of both the English and the Swedish data. In particular, had we applied the ideal observers to the exposure and test stimuli from both experiments *prior to collecting data*, we would have correctly predicted an effect for the English experiment and a null effect for the Swedish experiment. In this sense then, the Swedish experiment would *not* constitute a Type II error. The quality of fit was also confirmed by trial-level linear mixed-effects regressions reported in the SI (§5). These analyses found that the posterior probability of the /d/ category was a significant predictor of listeners' /d/-goodness ratings ($\hat{\beta} = 1.25$, $p < .001$). This effect remained significant when the experiment (English vs. Swedish), exposure group (/d/-exposure vs. control), and their interaction were included in the analysis ($\hat{\beta} = 1.59$, $p < .02$; for additional details, see SI, §5.2).

To further elucidate the reason for the differences in the ideal observers' predictions for the two experiments, Figure 4 shows the ideal observers' predictions for each of the items participants heard during test, shown in a 3D cue space. A distributional learning framework predicts failure to observe evidence for adaptation if a) the non-native exposure stimuli provide misleading information about the non-native stimuli during test or b) if the distributions of cues in the non-native exposure stimuli do not differ much from native distributions. From the first two rows of Figure 4, it is apparent that the predicted null effect for the Swedish experiment is an example of case b): rather than the /d/-exposure model performing badly on the test items, both the control and the /d/-exposure model perform well on the

test items. The reason for this is also obvious: the realization of native and non-native /d/ did not differ much for the Swedish recordings (see also Figure 3). For the English recordings, on the other hand, the cue distributions for the Mandarin-accented /d/ stimuli differed starkly from those of the native-accented /d/ stimuli. Deviating from native pronunciations, the Mandarin-accented talker showed no distinction between /d/ and /t/ in vowel and closure duration but clear separation along the burst dimension. This gave listeners in the /d/-exposure group a clear learning advantage over the control exposure group.

## 4 DISCUSSION

Critical reviews of standard practices in the psychological sciences have called out the tendency to dismiss null results as uninformative (Franco et al., 2014). A welcome consequence of this is that it is now easier to publish null results, often as failures to replicate. This reduces the "file drawer" problem (Rosenthal, 1979). The present work can be seen as building on this idea, aiming to understand *why* a null effect is observed. Specifically, the motivation for the present report grew out of an attempt to extend a previously replicated result of accent adaptation to a new L1-L2 pair, Flemish-accented Swedish. Apart from the language, test talker, and lexical materials, this experiment closely followed the design and procedure of previous work, specifically an experiment on Mandarin-accented English (Xie et al., 2017). Beyond the rating results from Xie and colleagues, several other studies with similar design had previously found the predicted effect of /d/-exposure, indexed either by increased auditory priming effects (Eisner et al., 2013; Xie and Myers, 2017; Xie et al., 2017) or improved segment identification (Xie et al., 2017). We thus expected that the experiment on Swedish would find positive evidence of adaptation, yet it seemingly failed to do so. After having ruled out differences in statistical power as a likely cause for the difference in results, we turned to computational models of speech perception to understand whether differences in the statistical properties of the exposure and test stimuli can explain the difference in results.

We found that ideal observers predict both the positive evidence for an effect for Mandarin-accented English in Xie et al. (2017) and the lack thereof in our experiment on Flemish-accented Swedish. This suggests that the original results were not a Type I error, nor are the Swedish results a Type II error. Rather, our ideal observer analyses suggest that the Swedish experiment would not find an effect even if repeated as a large-scale replication, at least as long as the same exposure and test stimuli are used. Indeed, even a much longer exposure phase that repeatedly presents the same non-native /d/ pronunciation as in our experiment on Swedish would not be expected to yield significant changes in participants' goodness ratings. The reason for this is clear from Figure 4: while the Flemish-accented talker differs from native speakers of Swedish in her realization of Swedish syllable-final /d/, these differences are small compared to the non-nativeness observed in the Mandarin-accented speech employed in the experiment on English.

At least qualitatively, ideal observer models provide a good fit against listeners' rating responses. This is remarkable since the modeling approach employed here does not include *any* degrees of freedom to mediate the effect of input statistics on perception. The only parameters of ideal observers describe the statistics of categories' cue distributions in the speech input. These parameters are thus not fitted to participants' responses during the perception experiment but rather to *production* data—specifically, speech data that is assumed to have formed listeners' prior expectations based on native speech input and speech data that listeners observe during exposure. Based on these speech data, ideal observers then make predictions about listeners' *perception* during a subsequent test phase (here goodness ratings). In this sense, ideal observers offer a particularly parsimonious explanation for the differences in results between the two experiments. Taken together with other findings, this lends support to the hypothesis that adaptation to non-native accents involves similar mechanisms as adaptation to more subtle differences between native

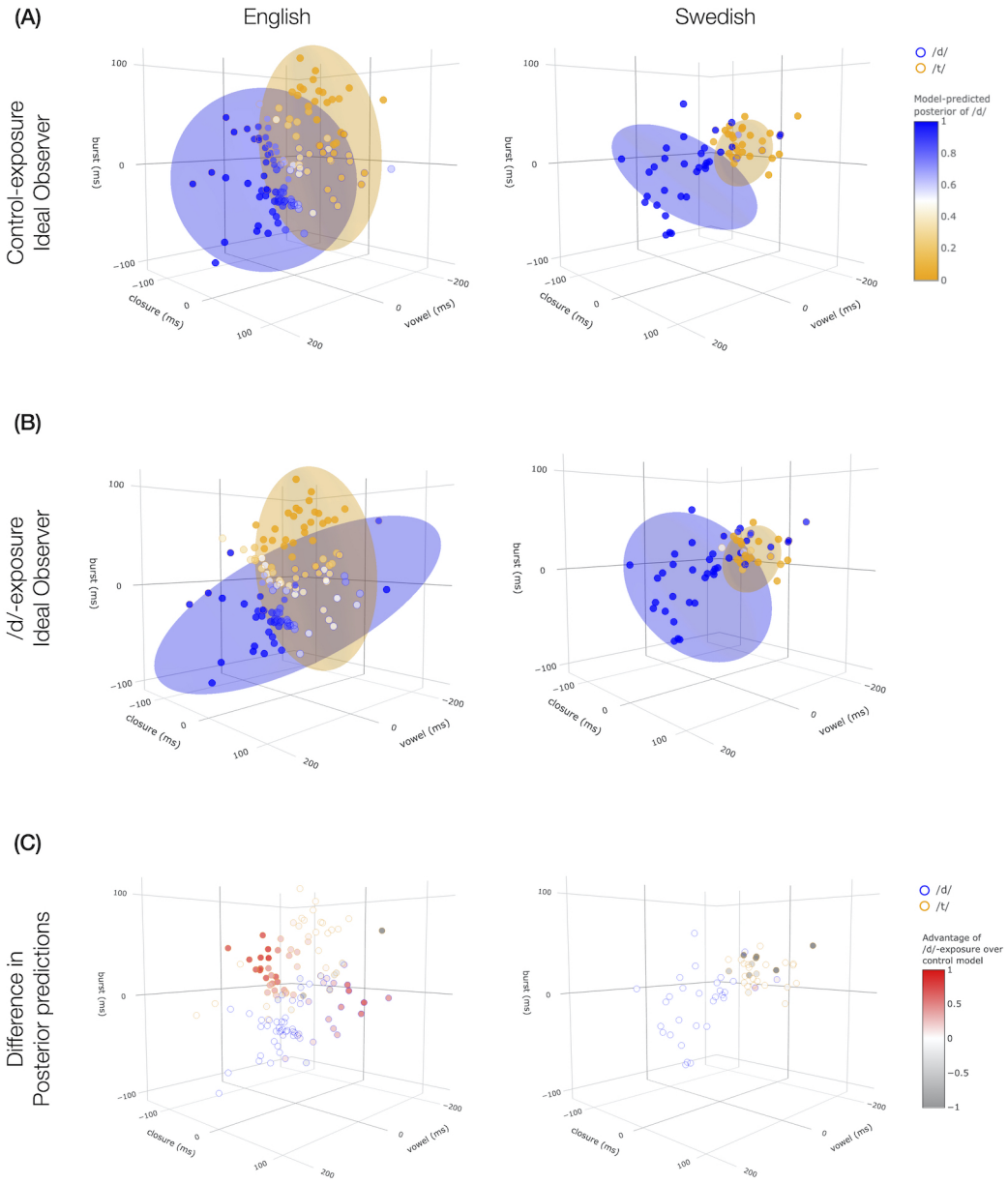


**Figure 4.** Predicted ratings of control (Panel A) and /d/-exposure ideal observers (Panel B), as well as their difference (Panel C) shown for all test tokens. These models were constructed from the distributions shown in §4.3.2 in the SI so as to simulate a learner who has either been exposed to, and perfectly learned, the statistics of the non-native /d/ (/d/-exposure model) or has not been exposed to non-native /d/ and thus assumes native /d/ statistics (control model). Both models assume native /t/ statistics (/t/ was never experienced in either exposure condition). Ellipses visualize the category likelihoods assumed by the respective ideal observers (specifically 95% of the probability density). Across all panels, the outline color of points indicate their intended category. In Panel A and B, the degree of color match between a point's outline color and fill color indicates a more accurate prediction matching the intended category. In Panel C, The color fill of points (bottom row) indicate the difference in posterior predictions between the /d/-exposure and control exposure models: redness indicates better performance in the /d/-exposure model (relative to control) and greyness indicates the opposite pattern. See SI:§5.2.1).

talkers (see also Eisner et al., 2013; Reinisch and Holt, 2014), and that these mechanisms include some form of distributional learning (see also Wade et al., 2007; Xie et al., 2017; Xie and Myers, 2017).

The present study also contributes to similar efforts to facilitate the theoretical interpretation of perception experiments through computational modeling (e.g. Chodroff and Wilson, 2018; Clayards et al., 2008; Feldman et al., 2009; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; McMurray and Jongman, 2011; Toscano and McMurray, 2010). In particular, an emerging body of work has used ideal observers and ideal adaptors to quantify how changes in the distributional statistics of phonetic cues affect listeners' categorization decisions (e.g., Clayards et al., 2008; Kleinschmidt et al., 2015; Kleinschmidt and Jaeger, 2016, 2011; Kleinschmidt et al., 2012; Theodore and Monto, 2019). When listeners are exposed to speech in which categories' cue distributions deviate from those of typical talkers—e.g., in terms of changes in categories' means or variances—this affects how listeners perceive and categorize subsequent input from the same talker. This manifests in changes in the location (Kleinschmidt and Jaeger, 2011; Kleinschmidt et al., 2012; Kleinschmidt and Jaeger, 2015) or the steepness of listeners' categorization functions (Clayards et al., 2008; Theodore and Monto, 2019) that are well-described by ideal observer and adaptor models. More recent work has begun to go one step further, using exposure-induced changes in categorization behavior from multiple exposure conditions to probe the structure of listeners' prior expectations about cross-talker variability (Kleinschmidt, 2020; Kleinschmidt and Jaeger, 2016).

Of direct relevance to the present study is recent work by Hitczenko and Feldman (2016). Like the present work, Hitczenko and Feldman employed computational models post-hoc to inform the theoretical interpretation of a previously reported finding from an experiment on adaptation to a synthesized accent (Maye et al., 2008). Maye and colleagues exposed listeners to synthesized American English in which all front vowels were simulated to have undergone phonological lowering (e.g., [i] became [ɪ] and [ɪ] became [ɛ], etc.). Listeners subsequently completed a lexical decision task of previously unheard words by the same synthesized voice with front vowels either lowered or raised. Based on the specific pattern of results, Maye and colleagues concluded that listeners adapted to the synthesized accent by shifting the means of their category representations, rather than merely becoming more accepting of *any* type of input. This finding and its interpretation has been influential, with almost 300 citations since 2008. Hitczenko and Feldman (2016) revisit these results, comparing them to the predictions of different types of ideal distributional learners (ideal adaptors, an extension to the simpler ideal observers employed here Kleinschmidt and Jaeger, 2015). Based on these computational comparisons, Hitczenko and Feldman conclude that shifted category representations are *not* the only way, or even the best, way to explain the specific changes in listeners' perception after exposure to the synthesized accent.

### 4.1 Limitations and future directions

These studies and the present work serve as examples of how fully specified computational models can inform theoretical interpretations of empirical findings. In our experience, many of the challenges in deriving principled predictions for experimental designs only become apparent during the process of implementing a computational model. To illustrate this we refer the reader to Table S1 in the supplementary material, which lists all of the assumptions we made in the present study. In the remainder, we discuss some of these assumptions, their limitations, and how future work might go about relaxing and revising them.

First, we made simplifying assumptions about what sources of noise contribute to listeners' estimates of the relevant cue distributions. Acoustic noise in the environment and neural noise in listeners' perceptual systems distort the speech signal produced by talkers beyond whatever variability results from noise during the planning and execution of speech articulation. By estimating distributions from speech recordings, our ideal observers ignore whatever acoustic noise our participants experienced beyond those in the recordings,

as well as any noise within listeners' perceptual systems.[2] This might explain why the responses predicted by the ideal observers are more categorical than the actual responses made by human listeners: adding perceptual noise to our ideal observers would increase the variance of cue distributions, leading to more shallow categorization functions, and thus less categorical predicted rating responses. Previous work has demonstrated that noise effects can be quantitatively estimated from separate perceptual data and integrated into ideal observers (Kronrod et al., 2016; Feldman et al., 2009). It would be informative to see whether the inclusion of perceptual noise improves the fit between the ideal observers' predictions and human perceptual decisions.

Second, for the present project, we applied normalization procedures on the acoustic cues to correct for phonotactic context effects. We made the simple assumption that such correction is based on native experience, regardless which accent listeners heard. That is, the /d/-exposure model assumed no learning of non-native phonotactic regularities. On the one hand, this would seem to be in the spirit of C-CuRE and related normalization approaches (McMurray and Jongman, 2011; Lobanov, 1971; Nearey, 1978). For example, C-CuRE computes acoustic cues relative to expectations about the mean of cues in a particular phonotactic or talker context. Critically, the C-CuRE model presented in McMurray and Jongman (2011) assumes that these adjustments are made independent of each other—i.e., this normalization procedure corrects for talker-specific differences in cue distributions and for phonotactics, but not for talker-specific phonotactics. On the other hand, there is evidence that non-native speech deviates from native speech in not only the overall realization of categories, but also in how specific phonotactic contexts affect pronunciation (as found in, e.g., Xie and Jaeger, 2020; Flege and Wang, 1989; Lahiri and Marslen-Wilson, 1991). Whether listeners in the accent adaptation experiments learn these non-native phonotactics in addition to changes in category-to-cue distributions is an open question. Future work could therefore compare models like ours without learning talker- or accent-specific phonotactic patterns against models that also learn this information.

Third, we constructed the /d/-exposure and the control models directly from the input statistics in each accent (non-native vs. native). These models assumed complete learning whereby listeners are assumed to have fully converged towards exposure statistics. In reality, rational listeners are expected to be guided by prior beliefs based on their native experience. While such priors facilitate adaptation to talker-specific statistics that meet prior expectations (Kleinschmidt and Jaeger, 2015), the same priors slow-down and constrain learning of unexpected non-native statistics (Kleinschmidt and Jaeger, 2016; Kleinschmidt, 2020). Learners are thus not expected to fully converge against the statistics experienced during exposure. Future work might consider the same type of incremental Bayesian belief updating applied in previous work on the perception of native speech (Kleinschmidt and Jaeger, 2011; Theodore and Monto, 2019) or synthesized speech (Hitczenko and Feldman, 2016) to investigate adaptation to the perception of non-native speech.

Beyond the aforementioned specifics of the models, there are limitations to the specific way in which the present study employed ideal observers: our approach has been both post-hoc and confirmatory. With regard to the latter, future work could follow in the footsteps of Hitczenko and Feldman (2016), and compare the ideal observers developed here against alternative hypotheses. For example, instead of distributional learning, the effects of different exposure on listeners' rating responses during test might reflect changes in response biases (Clarke-Davidson et al., 2008) or a general relaxation of response criteria (Hitczenko and Feldman, 2016). Similarly, future work might employ the same methods we have used here predictively. As we have illustrated in the present study, the distributional statistics of the specific input—and more

[2] At the same time, our ideal observers' estimates of all relevant cue distributions are likely perturbed by measurement errors due to the annotation procedure we used.

specifically the way in which such statistics differ between native and non-native speech—can be linked to predicted changes in subsequent perception. Future work can use ideal observer-predicted categorization or rating responses in power analyses to inform experimental designs prior to the experiment (for similar approaches in other domains, see Bicknell et al., under review; Jaeger et al., 2019).

On a related note and final note, the evidence that statistical distributions of speech cues have direct consequences on perceptual outcomes prompts us to reconsider intuitions that speakers within a given sociolect, dialect or in this case non-native accent group are equally easy or difficult to adapt to for a native listener of a particular language. By examining the cue-to-category mapping for particular talkers, we can predict to some extent, how listeners are able to accurately infer a target talker's intended speech as they converge towards the talker's distributional statistics. This points to opportunities for future work in studying how talker-specific acoustic distributions affect the degree and pace of adaptation.

## REFERENCES


Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance* 35, 520. doi:https://doi.org/10.1037/a0013552

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2021). *rmarkdown: Dynamic Documents for R*. R package version 2.7

Allen, J. S. and Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics* 63, 798–810. doi:https://doi.org/10.3758/BF03194439

Apfelbaum, K. S. and McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review* 22, 916–943. doi:https://doi.org/10.3758/s13423-014-0783-2

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

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience* , 171–206

Bicknell, K., Bushong, W., Tanenhaus, M. K., and Jaeger, T. F. (under review). Listeners can maintain and rationally update uncertainty about prior words

Booij, G. (1999). *The Phonology of Dutch* (Oxford University Press)

Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi:https://doi.org/10.1016/j.cognition.2007.04.005

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

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard* 4. doi:10.1515/lingvan-2017-0047

Clarke, C. M. and Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America* 116, 3647–3658. doi:https://doi.org/10.1121/1.1815131

Clarke-Davidson, C. M., Luce, P. A., and Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & psychophysics* 70, 604–618. doi:https://doi.org/10.3758/PP.70.4.604

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804–809. doi:https://doi.org/10.1016/j.cognition.2008.04.004

Drouin, J. R., Theodore, R. M., and Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America* 140, EL307–EL313. doi:https://doi.org/10.1121/1.4964468

Eisner, F. and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67, 224–238. doi:https://doi.org/10.3758/BF03206487

Eisner, F., Melinger, A., and Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology* 4, 148. doi:https://doi.org/10.3389/fpsyg.2013.00148

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review* 116, 752. doi:https://doi.org/10.1037/a0017196

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* 92, 233–277

Flege, J. E., Munro, M. J., and Skelton, L. (1992). Production of the word-final english/t/–/d/contrast by native speakers of english, mandarin, and spanish. *The Journal of the Acoustical Society of America* 92, 128–143. doi:https://doi.org/10.1121/1.404278

Flege, J. E. and Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/contrast. *Journal of Phonetics* 17, 299–315. doi:https://doi.org/10.1016/S0095-4470(19)30446-2

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science* 345, 1502–1505. doi:10.1126/science.1255484

Goslin, J., Duffy, H., and Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and language* 122, 92–102. doi:https://doi.org/10.1016/j.bandl.2012.04.017

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (Routledge). 249–307

Hitczenko, K. and Feldman, N. H. (2016). Modeling adaptation to a novel accent. In *Proceedings of the Annual Conference of the Cognitive Science Society*. 1367–1372

Idemaru, K. and Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance* 37, 1939. doi:https://doi.org/10.1037/a0025641

[Dataset] Jaeger, T., Burchill, Z., and Bushong, W. (2019). Strong evidence for expectation adaptation during language understanding, not a replication failure. a reply to Harrington Stack, James, and Watson (2018)

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics* 57, 21–39. doi:https://doi.org/10.1016/j.wocn.2016.05.001

Kleinschmidt, D. (2020). What constrains distributional learning in adults?

Kleinschmidt, D. and Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. 10–19

Kleinschmidt, D., Raizada, R., and Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci15)* (Austin, TX: Cognitive Science Society)

Kleinschmidt, D. F., Fine, A. B., and Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the annual meeting of the cognitive science society*. vol. 34

Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122, 148. doi:https://doi.org/10.1037/a0038695

Kleinschmidt, D. F. and Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? In *CogSci*

Kraljic, T. and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review* 13, 262–268. doi:https://doi.org/10.3758/BF03193841

Kraljic, T. and Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56, 1–15. doi:https://doi.org/10.1016/j.jml.2006.07.010

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review* 23, 1681–1712. doi:https://doi.org/10.3758/s13423-016-1049-y

Lahiri, A. and Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition* 38, 245–294. doi:https://doi.org/10.1016/0010-0277(91)90008-R

Lancia, L. and Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Laboratory Phonology* 4, 221–257. doi:https://doi.org/10.1515/lp-2013-0009

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49, 606–608. doi:https://doi.org/10.1121/1.1912396

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19, 1

Massaro, D. W. and Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review* 97, 225. doi:https://doi.org/10.1037/0033-295X.97.2.225

Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science* 32, 543–562. doi:https://doi.org/10.1080/03640210802035357

McMurray, B. and 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, 219. doi:https://doi.org/10.1037/a0022325

Munro, M. J. and Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech* 38, 289–306. doi:https://doi.org/10.1177/002383099503800305

Nearey, T. (1978). Phonetic feature systems for vowels (indiana university linguistics club, bloomington, in)

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., and Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: eye movement evidence from cantonese segment and tone perception. *Journal of Memory and Language* 90, 103–125. doi:https://doi.org/10.1016/j.jml.2016.03.005

Norris, D. and McQueen, J. M. (2008). Shortlist b: a bayesian model of continuous speech recognition. *Psychological Review* 115, 357. doi:https://doi.org/10.1037/0033-295X.115.2.357

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

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language* 45, 137–158

Porretta, V., Tremblay, A., and Bolger, P. (2017). Got experience? PMN amplitudes to foreign-accented speech modulated by listener experience. *Journal of Neurolinguistics* 44, 54–67. doi:https://doi.org/10.

1016/j.jneuroling.2017.03.002
Qian, T., Jaeger, T. F., and Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns. *Cognition* 157, 156–173. doi:https://doi.org/10.1016/j.cognition.2016.09.002
Reinisch, E. and Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance* 40, 539. doi:https://doi.org/10.1037/a0034409
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin* 86, 638. doi:https://doi.org/10.1037/0033-2909.86.3.638
Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics* 31, 307–314. doi:https://doi.org/10.3758/BF03202653
Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of phonetics* 52, 183–204. doi:https://doi.org/10.1016/j.wocn.2015.07.003
Schmale, R., Cristia, A., and Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science* 15, 732–738. doi:https://doi.org/10.1111/j.1467-7687.2012.01175.x
Shi, L., Griffiths, T. L., Feldman, N. H., and Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review* 17, 443–464. doi:https://doi.org/10.3758/PBR.17.4.443
Sidaras, S. K., Alexander, J. E., and Nygaard, L. C. (2009). Perceptual learning of systematic variation in spanish-accented speech. *The Journal of the Acoustical Society of America* 125, 3306–3316. doi:https://doi.org/10.1121/1.3101452
Smith, B. L., Johnson, E., and Hayes-Harb, R. (2019). ESL learners' intra-speaker variability in producing American English tense and lax vowels. *Journal of Second Language Pronunciation* 5, 139–164. doi:https://doi.org/10.1075/jslp.15050.smi
Tan, M. S. L., Xie, X., and Jaeger, T. F. (2019). Analysing L2 swedish word-final stops. In *10th Tutorial and Research Workshop on Experimental Linguistics*. 193–196
Theodore, R. M. and Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review* 26, 985–992. doi:https://doi.org/10.3758/s13423-018-1551-5
Toscano, J. C. and McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science* 34, 434–464. doi:10.1111/j.1551-6709.2009.01077.x
Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. *Phonetica* 76, 327–358. doi:https://doi.org/10.1159/000487269
Verhoeven, J. (2005). Belgian standard dutch. *Journal of the International Phonetic Association* 35, 243–247
Volaitis, L. E. and Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America* 92, 723–735. doi:https://doi.org/10.1121/1.403997
Wade, T., Jongman, A., and Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica* 64, 122–144. doi:https://doi.org/10.1159/000107913
Weil, S. (2001). Foreign accented speech: Encoding and generalization. *Journal of the Acoustical Society of America* 109, 2473
Witteman, M. J., Weber, A., and McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics* 75,

537–556. doi:https://doi.org/10.3758/s13414-012-0404-y

Xie, X., Buxó-Lugo, A., and Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition* 97. doi:https://doi.org/10.1016/j.cognition.2021.104619

Xie, X., Earle, F. S., and Myers, E. B. (2018a). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience* 33, 196–210. doi:https://doi.org/10.1080/23273798.2017.1369551

Xie, X. and Jaeger, T. F. (2020). Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America* 147, 3322–3347. doi:https://doi.org/10.1121/10.0001141

Xie, X. and Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language* 97, 30–46. doi:https://doi.org/10.1016/j.jml.2017.07.005

Xie, X., Theodore, R. M., and 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, 206. doi:https://doi.org/10.1037/xhp0000285

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., et al. (2018b). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America* 143, 2013–2031. doi:https://doi.org/10.1121/1.5027410

date pubstate

## CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## AUTHOR CONTRIBUTIONS

XX and MT conducted the behavioural experiments in English and Swedish respectively. TFJ and XX led in the statistical analyses; MT contributed to the statistical analyses. All three authors contributed to the writing of the manuscript and the supplementary information which is available online https://osf.io/72fkx/.

## FUNDING

This study was supported by a grant from the Knut & Alice Wallenberg Foundation (2018) awarded to MT.

## ACKNOWLEDGMENTS

We are grateful to Shawn Cummings for feedback on earlier versions of the manuscript.

## SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in OSF [https://osf.io/72fkx/].