Generalization in Accent Adaptation as Hierarchical Bayesian Inference

Jordan Hosier
Department of Linguistics
Northwestern University
Evanston, IL 60208
jhosier@u.northwestern.edu

Klinton Bicknell
Department of Linguistics
Northwestern University
Evanston, IL 60208
kbicknell@northwestern.edu

Abstract

With adequate experience, listeners improve their ability to comprehend accented speech. Previous work demonstrates that listeners who adapt to one accented talker generalize that adaptation to other accented talkers - exposure to multiple talkers of the same accent facilitates comprehension of a novel talker of that accent (Bradlow & Bent, 2008) and exposure to multiple novel accents facilitates comprehension of yet another novel accent (Baese-Berk, Bradlow, & Wright, 2013). We propose a model of accent adaptation in which the task of accent adaptation is represented as a problem of hierarchical Bayesian inference, which assumes that listeners simultaneously learn about the distribution of talkers, accent groups, and accented speech more generally. We show that a hierarchical Bayesian model can predict qualitative findings of generalization of accent adaptation, but that a non-hierarchical model cannot. The hierarchical Bayesian model also provides better quantitative predictions about sentence-level performance at test. Taken together, these findings support a view of accent adaptation as hierarchical inference.

1 Introduction

Effective speech recognition requires that listeners cope with variability, despite the associated computational difficulties. Work in perceptual learning demonstrates impressive coping strategies fundamental to speech processing despite the variability present (Norris, McQueen, & Cutler, 2003; Nygaard & Pisoni, 1998).

These coping strategies are also present when processing accented speech. Clarke and Garrett (2004) found that although initial processing speed is slower for accented speech, this deficit diminishes within one minute of exposure. For robust speech perception, listeners may learn about talkers and generalize what they have learned from one talker to another.

Sidaras, Alexander, and Nygaard (2009) found that listeners adapt to accent-general regularities and generalize those adaptations to novel accented talkers. Bradlow and Bent (2008) found similar generalization effects, providing evidence that accent adaptation occurs both talker-dependently and talker-independently. Beyond generalization of accent adaptation within and across talkers, there is also evidence of generalization across accents (Baese-Berk et al., 2013).

In this paper, we propose a model of generalization in accent adaptation as a form of hierarchical inference under uncertainty, where listeners simultaneously learn the properties of accented speech at talker-specific, accent-specific, and cross-accent levels. Previous work in has demonstrated that learners may benefit from exploiting underlying structure present in cross-situation variation to recognize familiar situations and generalize to similar situations (Pajak, Fine, Kleinschmidt, & Jaeger, 2016; Pajak, Bicknell, & Levy, 2013; Kleinschmidt & Jaeger, 2015; Nielsen & Wilson, 2008).

We use a new dataset of phonetically transcribed non-native-accented speech to determine if this model can reproduce the key empirical results known about generalization in accent adaptation and assess its quantitative ability to predict sentence-level comprehension of non-native-accented speech.

2 Generalization of Accent Adaptation

Much of the empirical evidence about generalization in accent adaptation comes from Bradlow and Bent (2008) and Baese-Berk et al. (2013). These studies had similar designs, in which participants were first exposed to a sequence of English sentences spoken in non-native accented speech from one or more talkers ('training'), and were tested on the speech of a Mandarin-accented talker and a Slovakian-accented talker ('test'). During both training and test, participants transcribed (in English) the foreign-accented speech, and the test portion of this was scored for transcription accuracy. The experiments manipulated which non-native accented talkers participants were exposed to during training. These studies found three critical differences between training conditions, which represent most of our empirical knowledge about how non-native accent adaptation is generalized across talkers and accents.

Talker Variability Advantage: Training on multiple talkers of an accent is more helpful when testing on that accent than training on a single talker of that accent. Bradlow and Bent (2008) demonstrated this principle by showing that participants trained on five Mandarin-accented talkers (Multi-talker training) performed better when tested on a new Mandarin-accented talker than participants trained on a single Mandarin-accented talker (Single-talker training).

Talker Specificity Advantage: Training on the same talker which you are later tested on is more helpful than training on multiple talkers of the same accent. Bradlow and Bent (2008) demonstrated this principle by showing that participants trained on the Mandarin-accented test talker (Talker-specific training) performed even better than participants with Multi-talker training.

Accent Variability Advantage: Training on talkers of multiple non-native accents is more helpful when testing on a new non-native accent than training on multiple talkers of a single non-native accent. Baese-Berk et al. (2013) demonstrated this principle by showing that participants trained on five talkers each with a different (non-Slovakian) non-native accent (Multi-accent training) performed better when tested on a Slovakian-accented talker than participants trained on five Mandarin-accented talkers.

3 Hierarchical Inference

Here, we propose that each of these findings about generalization in accent adaptation can be explained in terms of hierarchical inference. Specifically, we suggest that listeners learn about the speech patterns of individual talkers, abstract that to learn about regularities within non-native accent groups, and abstract that further to learn about regularities within non-native accents in general. Previous work has found robust evidence that hierarchical inference can well capture human language learning behavior (Pajak et al., 2016, 2013; Kleinschmidt & Jaeger, 2015; Hitczenko & Feldman, n.d.; Nielsen & Wilson, 2008).

Under this proposal, the Talker Variability Advantage would arise from the fact that multiple talkers of an accent give more information about properties of that accent than a single talker. Better information about the properties of that accent would yield better predictions for the properties of a new talker of that accent, and thus higher performance. For example, when being trained on just a single talker, it is unclear how much of that talker's speech patterns are idiosyncratic rather than representative of their accent group, whereas when being trained on multiple talkers of the same accent the commonalities are clearer.

The Talker Specificity Advantage, on the other hand, would arise in a different way in this proposal. Although multiple talkers of the same accent provide more information about the properties of an accent as a whole (and thus the best predictions about a new talker of that accent), a single talker provides the most information about that talker's own properties, including both idiosyncratic and accent-general elements.

Finally, the Accent Variability Advantage would have an explanation analogous to that of the Talker Variability Advantage. Just as multiple talkers of an accent provide the best information about that accent's properties that are useful to make predictions about a new talker of that accent, talkers of multiple non-native accents provide the best information about the properties of non-native accents in general and thus yield the best predictions about the speech of a talker of a new accent.

We test the extent to which hierarchical inference could be responsible for generalization of adaptation by presenting and evaluating a computational model of hierarchical Bayesian inference.

4 Model

In the current work, we characterize the properties of non-native speech in terms of segmental-level errors it contains (e.g., /θ/ → /z/), which have been associated with the most salient features of non-native-accented speech (Reinisch & Holt, 2014; Anderson-Hsieh, Johnson, & Koehler, 1992). The representation of errors at the segment-level allows for generalization across words while remaining specific to previously encountered segments (but see Linzen & Gallagher, in press, for evidence of rapid featural-level generalization).

We number types of segmental errors from 1 to n_e. We denote the number of times that talker i of accent k makes error j as $d_{i,j}^{(k)}$. Exposure to a training condition is formally denoted as observing a $d_{i,j}^{(k)}$ error count for each error type and each of $n_t(k)$ talkers in each of n_a accents k.

We formalize the task of accent adaptation as a form of Bayesian inference under uncertainty. Listeners are exposed to talkers that make errors a varying number of times and use this exposure to make predictions about how likely those talkers or some other talkers will be to make each error in the future. Our linking hypothesis to performance is that a listener with more accurate predictions about a talker's set of errors at test (i.e., a listener that assigns that talker's sets of errors a higher probability) will be more accurate at transcribing that talker's speech. We infer the probability via Bayesian inference in a formal generative model, which sets up a formal process by which accented speech is generated.

4.1 Hierarchical Bayesian model

The hierarchical Bayesian model infers three levels of hierarchical error rates: the error rates of individual talkers, the average error rates of distinct accents, and the average rates of errors across non-native accents. From these inferences, the model can make predictions about familiar talkers as well as novel talkers belonging to both novel and familiar accent groups. For these purposes, we assume that the model has perfect knowledge of talker identities and accent groupings. The assumption that listeners have perfect knowledge of when they have heard the same talker is a relatively weak assumption, in this case, as the only time listeners must recognize that the test talker is the same as the training talker is in the talker specific condition. In this condition, listeners hear the same talker across all training and test. We further assume here that all errors are independent of each other (observing the error /θ/ → /d/ doesn't affect the probability of making the error /θ/ → /z/). We deem this simplifying assumption to be a reasonable approximation of non-native speech errors. Relatedly, we assume that the errors talkers make are independent of the words they produce for computational simplicity. While this is empirically untrue, the more words talkers produce, the better approximation this will be.

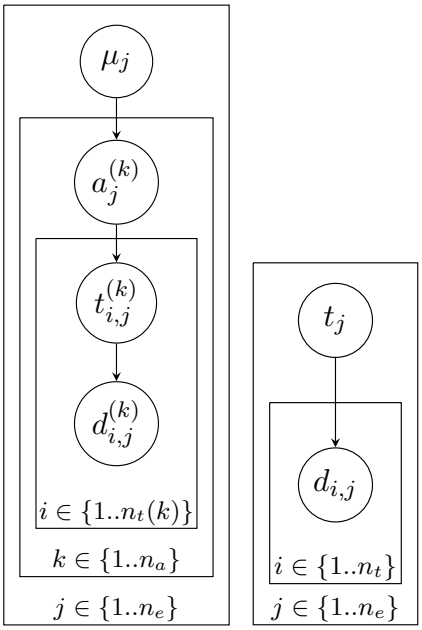

Figure 1: Graphical models for hierarchical Bayesian model (left) and baseline model (right).

4.1.1 Formal Generative Model

The highest level of hierarchy in the model (Figure 1 (left)) is the mean of (log) error rates across accents for an error type j, μ_j, which has a Gaussian prior with hyper-parameters μ_μ and σ_μ^2.

$$\mu_j \sim \mathcal{N}(\mu_\mu, \sigma_\mu^2) \tag{1}$$

Mean (log) error rates of each accent are normally distributed around this mean with variance σ_a^2.

$$a_j^{(k)} \sim \mathcal{N}(\mu_j, \sigma_a^2) \tag{2}$$

Within each accent, talkers' (log) error rates are drawn from a normal distribution centered at a mean $a^{(k)}$ with a variance of σ_t^2.

$$t_{i,j}^{(k)} \sim \mathcal{N}(a_j^{(k)}, \sigma_t^2) \tag{3}$$

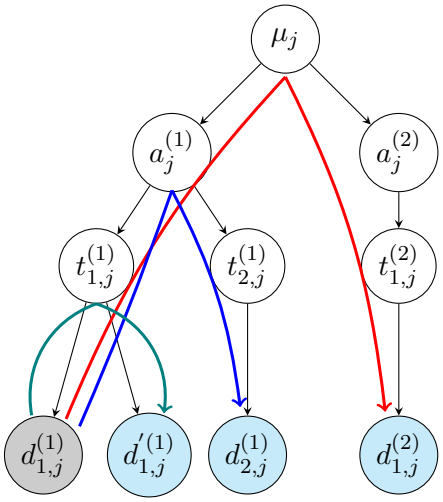

Figure 2: The three ways that a single observed $d_{i,j}^{(k)}$ node can help predict a new $d_{i,j}^{(k)}$ node. Within a talker (green) information flows from the observed d, to its respective t. Within an accent (blue) information flows from the observed d, to its respective t, up to that talker's a. Across accents (red) information flows from the observed d, to its respective t, up to that talker's a, further up to μ.

Finally, errors $d_{i,j}^{(k)}$ are Poisson-distributed given the rate for that accent, talker, and error type $t_{i,j}^{(k)}$ over that talker's amount of speech exposure, ℓ_i.

$$d_{i,j}^{(k)} \sim \text{Poisson}(\exp(t_{i,j}^{(k)}) \cdot \ell_i) \tag{4}$$

4.1.2 Inference

Bayesian inference inverts this graphical model. Exposure to error counts $d_{i,j}^{(k)}$ updates beliefs about that talker's rates of making that error $t_{i,j}^{(k)}$, which in turn updates beliefs about the average rates of making that error by talkers belonging to the same accent $a_j^{(k)}$, which updates beliefs about the overall rate at which non-native accents produce that error μ_j. This hierarchical structure allows the model to generalize strongly from a talker's errors at one time to that same talker's errors at another time, relatively weaker generalization from one talker to another within the same accent group, and the weakest generalization from a talker of one accent to a talker of a different accent group.

The model makes strong generalizations within a talker because listeners make direct inferences about the talker's error distribution from that talker's (log) error rates (Figure 2). Within an accent, the model makes relatively weaker generalizations. Listeners not only infer the first talker's error rates but also other talkers' of the same accent and subsequently use the distribution of accent-level (log) error rates to make predictions about what a new talker's error rates will be. Nonetheless, increasing the amount of exposure to talkers belonging to one accent group allows the model to be more confident in its predictions about error space for that accent. Across accent generalization is the weakest as information flows further up the tree than in the previous two cases. Listeners simultaneously infer multiple talker (log) error rates belonging to different accent groups and use the distribution of non-native error rates to make inferences about a novel talker and accent.

4.2 Baseline model

To determine the extent to which *hierarchical* inference is critical to the predictions of the model, we also construct a Baseline model that works similarly, but lacks hierarchy, as seen in Figure 1 (right). This model cannot distinguish between different talkers or accents. Thus, the generalization process is equivalent following exposure to the same talker at training and test, differing talkers of the same accent, and different accents.

4.2.1 Formal Generative Model

Errors are taken to be Poisson-distributed given an overall log rate parameter μ_j.

$$d_{i,j} \sim \text{Poisson}(\exp(\mu_j) \cdot \ell_i) \tag{5}$$

Overall (log) error rates have the same prior as the overall (log) rates in the hierarchical model.

$$\mu_j \sim \mathcal{N}(\mu_\mu, \sigma^2) \tag{6}$$

4.2.2 Inference

There is only one path of inference in this model (Figure 3): listeners infer the overall rate for each error and use this rate to predict the error data of a test talker, regardless of any relationships between test and training talkers. Thus, inferences are equally strong within a talker as across accents.

5 Qualitative Evaluation

To evaluate model performance, we test its ability to predict the qualitative patterns seen in Bradlow and Bent (2008) and Baese-Berk et al. (2013). Specifically, we use the model to compute, for

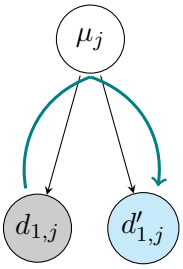

Figure 3: Graphical representation of the flow of information through the baseline model.

each condition, the probability of the errors made by a test talker given the errors made by the training talkers. Here, we aim to see if the model can capture the three advantages seen in prior work on generalization in accent adaptation (Bradlow & Bent, 2008; Baese-Berk et al., 2013). The model is deemed successful to the extent that it predicts the test talker to be higher probability in the training condition that saw better performance.

5.1 Method

5.1.1 Dataset

The non-native speech from both training and test talkers in the Bradlow and Bent (2008) and Baese-Berk et al. (2013) studies was transcribed to IPA by two phonetic-trained native-English transcribers. Any differences were resolved by a third such transcriber.

5.1.2 Alignment

For each word that did not match its correct pronunciation, we estimated the best alignment of the produced and correct IPA pronunciation using the Needleman-Wunsch Algorithm (Needleman & Wunsch, 1970), which finds the minimal set of insertions, deletions, and substitutions needed to transform the correct pronunciation into the produced pronunciation.[1] From these alignments, we can extract the implied set of segmental phonological errors that were made. For example, if /fɑdə/ was uttered instead of the intended /fɑðər/, the two errors would be /ð/ → /d/ and /r/ → ∅.

The resulting dataset, comprised of all test and training talkers, contained 265 error types with a total of 1,155 errors made.

5.1.3 Generating Model Predictions

We set out to test the three qualitative properties of generalization in accent adaptation established in Bradlow and Bent (2008) and Baese-Berk et al. (2013). Each qualitative property represents a significant difference between two training conditions for a particular test talker. The models displayed in each of the figures show the path of inference from a training condition to every possible test condition because participants were tested on both a Mandarin-accented speaker as well as a Slovakian-accented speaker.

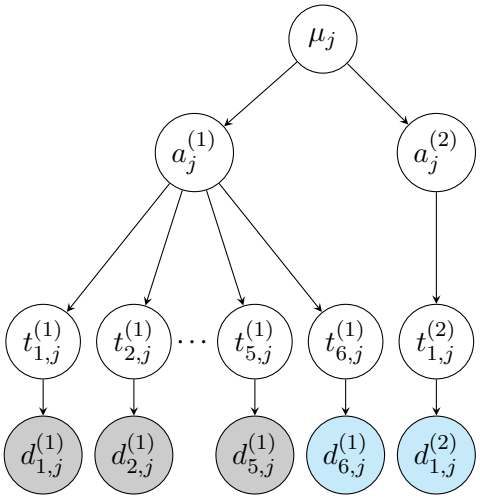

Figure 4: Graphical model of Multi-talker training with observed training data in gray and predicted test data in blue.

Talker Variability Advantage: To test the Talker Variability Advantage, we compare Multi-talker training to Single-talker training and generate predictions about a Mandarin test talker in each. The Multi-talker model (Figure 4) takes as input error data from five Mandarin-accented talkers in the Multi-talker condition $d_{i,j}^{(1)}$ where $i \in \{1...5\}$. The model makes predictions about a novel Mandarin-accented talker, $d_{6,j}^{(1)}$.

The Single-talker model (Figure 5) observes data from a single Mandarin-accented talker $d_{1,j}^{(1)}$ and makes predictions about a novel Mandarin-accented talker, $d_{2,j}^{(1)}$. There were four different talkers used as training talkers in this condition. Their resulting log probabilities were generated separately and averaged to compute the overall Single-talker training probability.

Talker Specificity Advantage: To test the Talker Specificity Advantage, we generate model predictions for the Multi-talker condition, explained above, and the Talker-specific condition (Figure 6). This model was initialized using the data from the Talker-specific condition which was

[1] For ties, we preferred substitution then deletion.

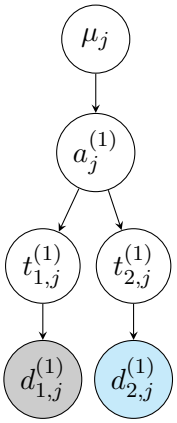

Figure 5: Graphical model of Single-talker training and Mandarin test where observed training data is in gray and predicted test data is in blue.

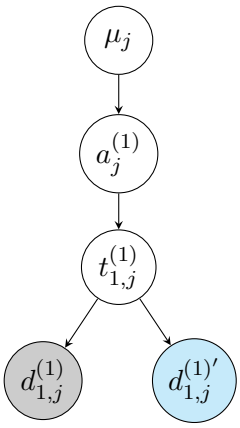

Figure 6: Graphical model of Talker-specific training and Mandarin test with observed training data in gray and predicted test data in blue.

observed by the model in the form of $d_{1,j}^{(1)}$. Thus, $t_{1,j}^{(1)}$ is being inferred from $d_{1,j}^{(1)}$ and used to generate predictions about the same Mandarin-accented talker $d_{1,j}^{(1)'}$.

Accent Variability Advantage: Finally, to test the Accent Variability Advantage, we generate model data for the Multi-talker condition and Multi-accent with Slovakian test. The Multi-talker model (Figure 4) was initialized with the same training data as in the Mandarin test case, but predictions were generated for a new talker with a novel accent $d_{1,j}^{(6)}$. The Multi-accent model (Figure 7) was initializedthe model predictions improve. When the parametrization has a higher value of using training data from the Multi-accent condition in the form of $d_{i,j}^{(k)}$ where $i \in \{1...5\}$ and $(k) \in \{1...5\}$. Predictions are made for new talker

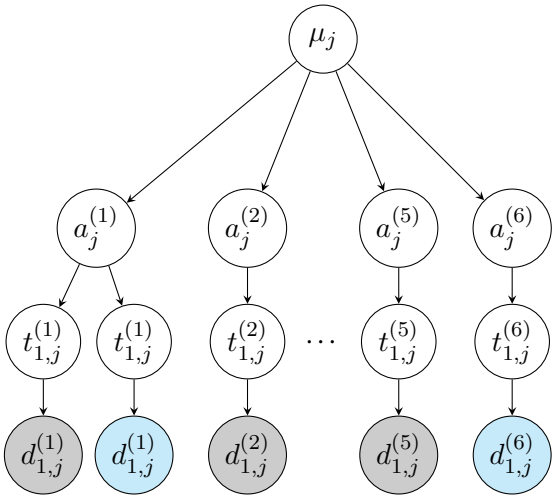

Figure 7: Graphical model of Multi-accent training with observed training data in gray and predicted test data in blue.

of a new accent $d_{1,j}^{(6)}$.

5.1.4 Inference

We estimate the probabilities of the errors made by a test talker given the model samples and a particular set of errors observed in training in a two-step process. As a first step, we performed Hamiltonian Markov chain Monte Carlo (MCMC) sampling on all unknown variables in the model[2], but excluding the $d_{i,j}^{(k)}$ representing the test talker's errors. We performed MCMC in RStan with 5,000 samples for each of 4 chains, and excluded the first half of each chain as warm-up, resulting in 10,000 MCMC samples. To assess convergence, we verified that the $\hat{R}$ of each sampled variable was less than 1.1.

As a second step, we computed the probability of the errors made by the test talker by averaging the probabilities of these errors under each of the 10,000 sampled $t_{i,j}^{(k)}$ values.

5.1.5 Parameters

We select variance parameters represented by σ_t^2 and σ_a^2 σ_μ^2. These variances could, in principle, be estimated using the error dataset. However the values could take on an infinite number of possible negative log error rates which would make for difficut estimation accuracy. For the hierarchical Bayesian model, variance parameters σ_t^2 and σ_a^2 vary between values of .5 and 4, representing large and small expectations of similarity within

[2]For computational simplicity, we integrated out $a_j^{(k)}$ and μ_j nodes where possible.

hierarchical Bayesian model

	σ_a^2 = .5	σ_a^2 = 4
σ_t^2 = .5	X	X
σ_t^2 = 4	✓	✓

baseline model

	σ^2 = 5	σ^2 = 12
	X	X

Table 1: Check marks encode the model's ability correctly predict the Talker Variability Advantage.

and across accents. The σ_μ^2 used in the prior on μ is set to 4.

The single variance parameter of the baseline model σ^2 is semantically equivalent to the sum of the σ_t^2, σ_a^2, and σ_μ^2 parameters from the hierarchical model. To use values at the extremes of the parameter space we tested for the hierarchical model, we thus test $\sigma^2 = 5$ and $\sigma^2 = 12$, which is derived from σ_t^2 + σ_a^2 + σ_μ^2 where σ_μ^2 is initialized at 4.

The hyper-prior mean μ_μ is initialized at $\log(.5402)$ in each instantiation of the model, computed as the log of the mean rate of errors seen in all training talkers. The ℓ_i values are set to the number of lists of sentences spoken by each training talker and test talker.

5.2 Results

We evaluate the model's performance by its ability to predict the three critical advantages present in previous work in accent adaptation. The model is deemed successful to the extent that it recovers each of these three advantages.

5.2.1 Talker Variability Advantage

As seen in Table 1 (top), performance of the hierarchical Bayesian model depends on the variance values used. When the across talker variation is low, the hierarchical Bayesian model cannot recover the Talker Variability Advantage. However, as σ_t^2 increases, the model predictions more closely resemble human performance.

With low between-talker variance (σ_t^2 = .5), the model is equally confident in a new talker's error distribution regardless of whether it was trained on a single talker or five distinct talkers. We see the model recover the Talker Variability Advantage as the between-talker variance increases because the model needs many independent data points from that accent (i.e. different talkers) to be confident in its predictions. Increasing the variance parameters decreases the amount of generalization because the model expects talkers and accents to be more different from one another.

The baseline model lacks the hierarchy necessary to learn about talker-specific, accent-specific and accent-general properties. Its generalizations from talker to talker are equally strong regardless of training condition. The baseline model predicts the test data to be higher probability given training in the Single-talker condition than the Multi-talker condition (Table 1, bottom). Because the baseline model is unable to differentiate talkers, exposure to many talkers in the Multi-talker condition results in false confidence about what a new talker's error distribution will look like, making performance in this condition worse than that of the Single-talker condition.

5.2.2 Talker Specificity Advantage

Both the hierarchical Bayesian model and the baseline model reveal the effects of the Talker Specificity Advantage regardless of the variance parameters selected. This result is reasonable given that this model makes predictions based on talker similarity. When the training talker is the same as the test talker there is a great deal of similarity in the distribution of errors making this condition higher probability. Further, we see the same disadvantage of overconfidence in the baseline model for the Multi-talker condition.

5.2.3 Accent Variability Advantage

Both the hierarchical Bayesian model and the baseline model exhibit the Accent Variability Advantage regardless of the variance parameters selected. The structure of the hierarchical model allows for different inferences to be made for distinct accents. The baseline model, because it cannot differentiate between talkers, should, in principle, see equivalent performance in Multi-talker and Multi-accent training. However, the baseline model captures the Accent Variability Advantage because empirically the error distribution of the Slovakian test talker is more similar to the Multi-accent condition.

5.3 Discussion

By formalizing the task of accent adaptation as a form of Bayesian inference, the baseline model successfully recovers the Talker Specificity Advantage as well as the Accent Variability Advan-

tage. The Hierarchical Bayesian model successfully recovers both these advantages as well as the Talker Variability Advantage. The hierarchical captures the ability of speakers to do narrow generalization - generalization across individuals yields an accurate accent-level hypothesis.

6 Quantitative Evaluation

The evaluations of the previous section revealed that the hierarchical Bayesian model predicts qualitative patterns seen in previous work where in one case the baseline model could not. However, that analysis had limited power. We would like to evaluate the model's predictive power at a more fine-grained level – performance at the sentence level. Here, we use mixed-effects regression models to analyze the models' ability to predict participant performance.

In Bradlow and Bent (2008) and Baese-Berk et al. (2013), participants were tasked with transcribing accented speech. We obtained the participant transcription data from Baese-Berk et al. (2013). The transcriptions were later scored at the sentence level. Here, we test the model's ability to predict sentence-level accuracy in the Baese-Berk et al. (2013) study.

6.1 Method

In the previous evaluation, we tested the model's ability to make correct predictions in the aggregate. The scores associated with each participant were based on the number of key words correct per sentence. The data for this evaluation includes only the Multi-talker and Multi-accent training conditions with Mandarin and Slovakian test talkers. We scored the participant data by whether the participant transcribed all key words correctly in the sentence.

6.1.1 Generating Model Predictions

We generate predictions for two training conditions, Multi-talker and Multi-accent, and for both a Slovakian-accented and a Mandarin-accented test talker. These conditions are visualized in Figure 4 and Figure 7. For each training condition, we estimate the log probability of the errors contained within each test sentence given the observed training data.

6.1.2 Parameters

We again test a number of different parameterizations of σ_a^2 and σ_t^2, once again testing values ranging between .5 and 4 for the hierarchical Bayesian model, but now exploring the space fully, σ_a^2 and $\sigma_t^2 \in \{0.5, 1, 2, 4\}$. The σ_μ^2 value is again set to 4. We cover a similar space for the baseline model, with $\sigma^2 \in \{5, 6, 8, 12\}$.

The hyper-prior mean μ_μ, is again initialized at $\log(.5402)$. The ℓ_i parameters are set as before for training talkers and to 1/16 of a list for each test talker, as we are making predictions at the sentence level and lists contained 16 sentences.

6.1.3 Analysis

We quantitatively assess the predictive power of these log probabilities for each sentence and condition in a logistic mixed-effects regression model predicting binary sentence accuracy. These regressions included a single fixed effect of interest, the model-generated log probability of the test sentence in the particular training condition, and also a fixed effect of the number of errors in this test sentence. Random intercepts for each participant and sentence were also included [3]

We assess the strength of model predictions in two ways. First, we compare each model to another similar logistic model that replaced the fixed effect of model-predicted log probability with a 4-level effect of experimental condition (2 training conditions crossed with 2 test conditions). Outperforming this baseline means that the model's predictions provide a better unified way of summarizing the data beyond knowing the main effects and interactions of training and test conditions. We formally compare these models using Bayes factors, which can be approximated by exponentiating half their difference in Bayesian information criterion (BIC) values (Wagenmakers, 2007). The strength of each Bayes Factor is visualized in Table 2 (Kass & Raftery, 1995). We further assess the strength of the hierarchical model predictions by comparing it to the logistic model with a fixed effect of the single best baseline model-predicted log probability.

6.2 Results

The results of this quantitative analysis are visualized in Figure 8 and Figure 9. The log probabilities estimated from the hierarchical Bayesian model vary in predictive quality across variance parameter space. In certain parts of the space, namely as σ_a^2 increases, the Bayes Factor com-

[3] Because the goal is not significance testing against a null hypothesis, we do not include random slopes, to simplify the model.

Strength of Bayes Factor

2 log B	B	Evidence against H_0
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
>10	>150	very strong

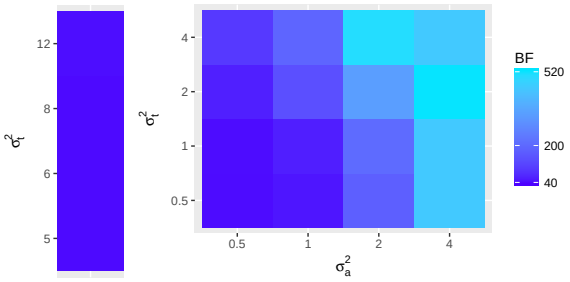

Figure 8: Bayes Factor comparing model predictions to condition labels for 4 variance parameter values in the baseline model (left) and 16 in the hierarchical Bayesian model (right).

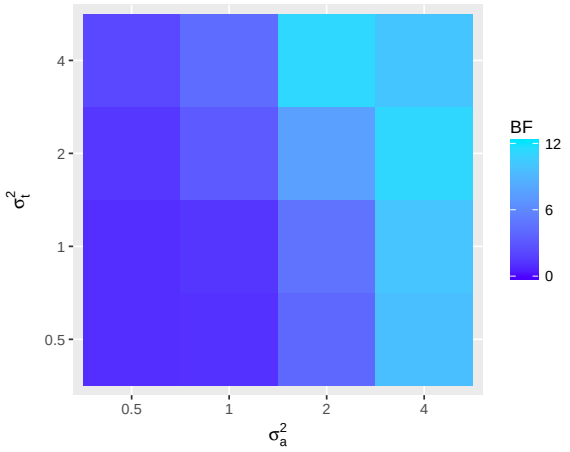

Figure 9: Bayes Factor comparing hierarchical model predictions to baseline model predictions.

paring model predictions as a predictor variable to condition as a predictor variable reaches very high values, demonstrating the strength of the hierarchical model's predictions. Predictions also generally improve as σ_t^2 increases. We saw the same effect in the qualitative evaluation: increasing the variances decreases the amount of generalization in that the model expects talkers and accents to be more distinct from one another.

Throughout variance parameter space, the baseline model's predictive power does not change and does not reach a Bayes Factor above 46 (Figure 8, left). Nevertheless, even a Bayes Factor of 46 represents "strong" evidence that the predictions of the baseline model are also to be preferred to simple condition labels.

When the two models are compared (Figure 9) we see a similar pattern across parameter space. Specifically, we yet again see positive evidence, in some parts of parameter space, that the hierarchical model is a better predictor of the test data than the baseline model.

6.3 Discussion

While both the baseline model and the hierarchical model are stronger predictors than training condition, the hierarchical model outperforms the baseline model. The hierarchy present in the hierarchical Bayesian model is crucial to its predictive power. The predictive power of the hierarchical Bayesian model increases as σ_a^2 increases. Regardless of the σ^2 parameters selected for the baseline model, its predictive power does not change.

7 General Discussion

The two models introduced in this work assume that the task of accent adaptation is a form of Bayesian inference. Crucially, however, the model which successfully recovers all of the three advantages of accent adaptation is the model which incorporates hierarchy into the modelling process. This model not only predicts these qualitative patterns, but it is a stronger predictor of sentence-level performance than training condition in the Baese-Berk et al. (2013) study. Taken together, these results support a view of accent adaptation as hierarchical Bayesian inference.

Acknowledgments

We thank Alexandra Saldan, Victoria Kuritza, and Michael Blasingame for their transcriptions of

(Bradlow & Bent, 2008) and (Baese-Berk et al., 2013). We also thank Bozena Pajak for her preliminary work with this project - first conceptualizing generalization of accent adaptation as a form of hierarchical inference as well as spearheading the development of the new dataset. We thank Ann Bradlow for her input and expertise. Finally, we thank our other colleagues in the Linguistics department for their feedback.

References

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language learning*, *42*(4), 529–555.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Hitczenko, K., & Feldman, N. H. (n.d.). Modeling adaptation to a novel accent.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

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.

Linzen, T., & Gallagher, G. (in press). Rapid generalization in phonotactic learning. *Laboratory Phonology*.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.

Nielsen, K., & Wilson, C. (2008). A hierarchical bayesian model of multi-level phonetic imitation. In *Proceedings of the 27th west coast conference on formal linguistics* (pp. 335–343).

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*, *60*(3), 355–376.

Pajak, B., Bicknell, K., & Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. In *Proceedings of the 4th workshop on cognitive modeling and computational linguistics* (pp. 11–20).

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: insights from first language processing. *Language Learning*, *66*(4), 900–944.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306–3316.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, *14*(5), 779–804.