"Oh, I've Heard That Before": Modelling Own-Dialect Bias After Perceptual Learning by Weighting Training Data

Rachael Tatman
Department of Linguistics
University of Washington
rctatman@uw.edu

Abstract

Human listeners are able to quickly and robustly adapt to new accents and do so by using information about speaker's identities. This paper will present experimental evidence that, even considering information about speaker's identities, listeners retain a strong bias towards the acoustics of their own dialect after dialect learning. Participants' behaviour was accurately mimicked by a classifier which was trained on more cases from the base dialect and fewer from the target dialect. This suggests that imbalanced training data may result in automatic speech recognition errors consistent with those of speakers from populations over-represented in the training data.

1 Introduction

Accent adaptation is an on-going challenge for automatic speech recognition (ASR) systems. However, a system does exist which adapts quickly and robustly to new accents: human speech perception, and in particular perceptual learning of new dialects.

This is a process which has been extensively researched in the fields of sociolinguistics and phonetics. It is well established that listeners do learn to perceive new language varieties (Fox, 1974; Hay et al., 2010; Sumner, 2011), and that they use world knowledge and their social beliefs about speakers to do so (Niedzielski, 1999; Johnson et al., 1999; Hay and Drager, 2010).

While modelling perceptual learning may lead to more human-like accent adaptation, to be a good cognitive model, it must capture behaviours found in experimental work: own-dialect bias, and the use of world knowledge during adaptation.

2 Experiment

2.1 Mainstream United States and New Zealand English

This experiment used acoustic data taken from sociologically-matched speakers of two distinct dialects: Mainstream US English (MUSE) and New Zealand English (NZE). The front vowels of MUSE and NZE vowels very confusable; due to an on-going vowel shift in NZE, the vowel in NZE "head" is very similar to the vowel in MUSE "hid", and NZE "had" sounds very like MUSE "head" (Watson, 2014). This overlap can be seen in Figure 3. Naive MUSE speakers listening to NZE, then, tend to make systematic errors when reporting which word they have heard. To evaluate this confusion, this experiment used a lexical identification task, where participants reported which word they believe a sound to be from.

2.2 Experimental Paradigm

First, listeners took a demographic questionnaire to ensure that they were from the United States, that they had not travelled to New Zealand and that they were unfamiliar with New Zealand English. Any listener not meeting these criteria was excluded from the experiment.

Listeners completed three lexical identification tasks. In each task, they were played a 300ms audio sample taken from the steady-state of the vowel from one of the words "heed", "head" or "had". Only the vowel was used in order to avoid possible confounds of variation in stop-production across dialects (Lyle, 2008). "Hid" tokens were excluded due to a very salient duration contrast in NZE between the vowel in "hid" and other vowels (Langstrof, 2009). Audio data was taken from two speakers of NZE and one speaker of MUSE. All speakers were white women between the ages of 20 and 30, with a college-level education or

higher.

The first lexical identification task was designed to train listeners to correctly identify the vowels of NZE. Listeners were played a vowel and asked to input which word it had come from: "heed", "head", "had" or "hid". If they picked the right word, they were told that it was correct and played a different token. If they picked the wrong one, they were told that they were incorrect, which word the token had been taken from and were given another chance to answer correctly. This process continued until each listener correctly labelled ten tokens in a row. For this task, listeners heard audio a single speaker of NZE and were told that the speaker was from New Zealand.

For the next two tasks, listeners did not receive feedback. In one task they were played audio from a MUSE speaker, and in the other from a second NZE speaker. They were explicitly told before each task the national origin of the speaker they were about to hear: New Zealand or the United States. However, while half of the listeners were given the correct information about each speaker, the other half were given incorrect information about each.

Twenty-one listeners participated in this experiment, most of whom (17) were from the West dialect region (Labov et al., 2005). The experiment was administered on-line using PsyToolkit (Stoet, 2010), and participants were recruited using social media and word of mouth.

3 Experimental Results

All participants were able to learn to successfully label the vowels of New Zealand English. This can be seen be seen in Figure 1: when given the correct social information about the speaker, the classifications of vowels from both MUSE and NZE were generally correct, with an average F_1 across classes of 0.96 and 0.75, respectively.

Being given incorrect information about a speaker's national origin, e.g. being told that a speaker from New Zealand was from the United States affected listeners classifications differently depending on the actual regional origin of the speaker. While listeners were still fairly accurate on their own dialect (F_1=0.81), the categorically changed their judgements for NZE (F_1=0.56). This difference can clearly be seen in Figure 2. The categorical changes are in line with confusions discussed in Section 2.1. NZE "head" is mis-classified as "hid", and NZE "had" is mis-classified as "head".

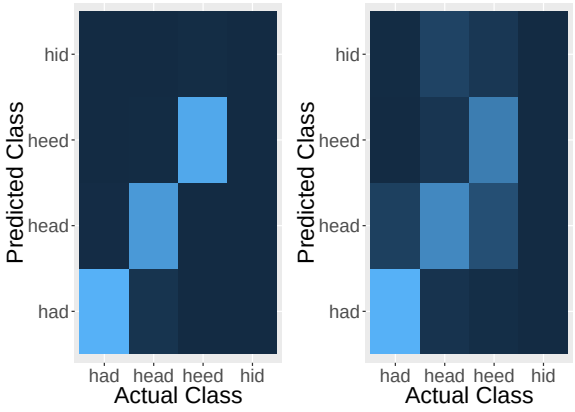

Figure 1: Heatmaps of the confusion matrix for MUSE (left) and NZE (right) classifications when participants were told the correct regional origin of the speaker. Note that, in both cases, most classifications were correct, i.e. on the center diagonal.

In other words, when listeners from the United States are listening to a speaker from New Zealand that they have been told is also from the United States, they are classifying NZE vowels as if they were from MUSE. However, the inverse is not true. Even when told that a MUSE speaker is from New Zealand, listeners are still classifying their vowels as if they believed them to be from the United States. This preference for vowel classifications in line with those of a listener's own dialect will be referred to as "own-dialect bias" from this point on.

4 Computational Modelling

Based on the experimental data outlined above, any classifier of multi-dialect data should have the following qualities in order to achieve human-like classifications:

- Classifications should depend on information about the speaker's dialect.
- The classifier should should also consider one dialect the "default" and be biased towards it.

4.1 Data

To maximize parallelism, the same acoustic data was used to train classifiers as the participants were given. For each item, four features were recorded: the speaker's regional origin (NZ or

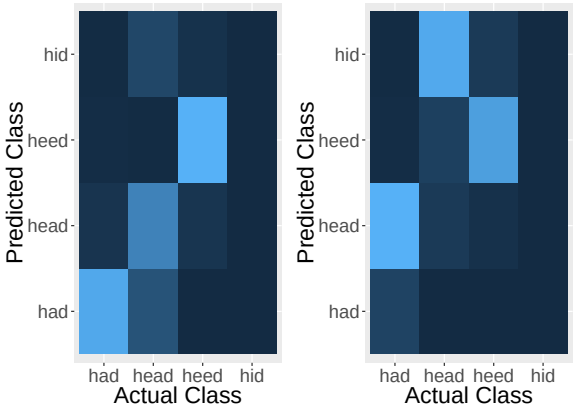

Figure 2: Heatmaps for the confusion matrix for MUSE (left) and NZE (right) classifications when participants were told the correct national origin of the speaker. While this had a slight effect on the classification of MUSE vowels, it categorically changed that of NZE vowels–this is what is meant by "own-dialect bias".

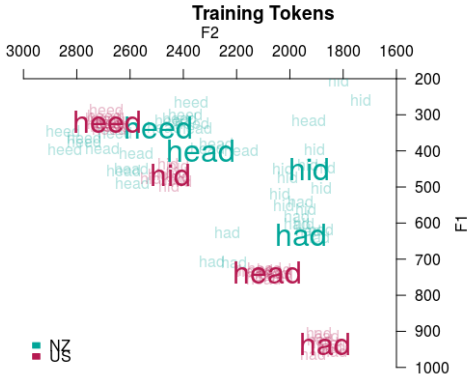

Figure 3: Plot of training tokens in a first by second formant space.

US), and the first, second and third formants as measured at the central point of the steady state of the vowel. Formants are frequency bands of high acoustic energy within the speech signal, and they are a robust cue to the identity of the vowel for human listeners (Wright, 2004). The vowels considered in this study are primarily distinguished by the first formant, as can be seen in Figure 3.

4.2 Conditional Inference Trees

Throughout this section, modelling is done using conditional inference trees (Hothorn et al., 2006). These are decision trees where nodes are split based on statistical inference rather than information gain. As a result, they are not biased towards factors with more levels, including numeric factors like formant values.

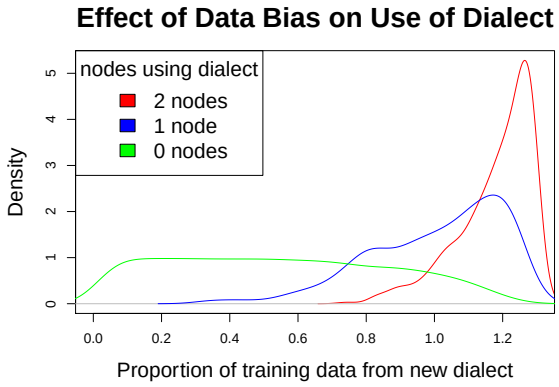

Figure 4: Figure showing how bias in training data affects the degree to which dialect information is used during classification. The more data from the second dialect is added to the original dialect, the more nodes in conditional inference tree will split based on dialect.

Another benefit of conditional inference trees is ease of interpretation. It is clear how each feature is being used during classification and to what extent it is affecting classifications, which allows a clear comparison between different conditional inference trees.

4.3 Bias as Exposure

One way to model own-dialect bias is to compare models trained on datasets where there is bias towards one dialect. To simulate this, 4500 conditional inference trees were trained. The proportion of NZE training tokens to MUSE token was varied between 0.02 (only 1 NZE token) to 1.3 (45 NZE tokens), with a uniform distribution. As the proportion of training tokens from the second dialect increased, so did the number of nodes in the conditional inference tree splitting based on speaker dialect. This can be seen in Figure 4.

A classifier trained on a dataset highly biased towards MUSE tokens was an excellent model of participants' own-dialect bias on mislabelled tokens from their own dialect. This can be seen in Figure 5. The classifications on the left are from a model trained on data strongly biased towards MUSE: 90% of the training data for this model came from MUSE. The classifications on the right are from a model trained on a dataset balanced between MUSE and NZE (50% each). The outputs of both models are based on MUSE data that has been mislabelled as being from NZE.

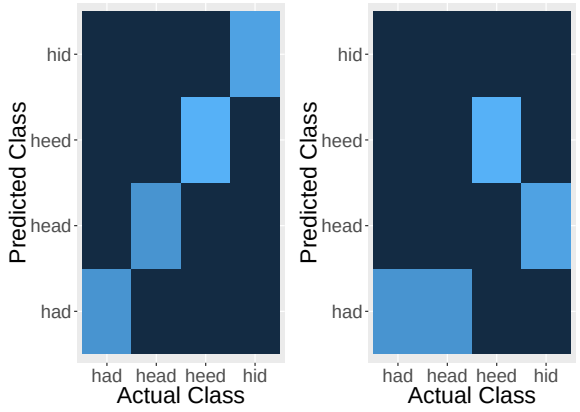

Figure 5: Classifications by conditional inference trees trained on data biased towards MUSE (on left) or balanced between MUSE and NZE (on right). These two classifiers are both showing classifications on MUSE tokens mislabelled as NZE. In this case, the biased classifier is a better behavioural model.

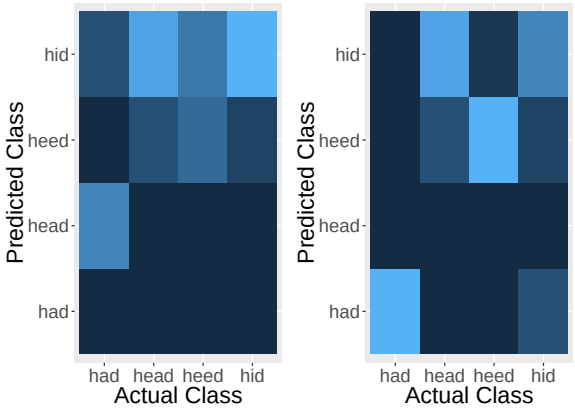

Figure 6: Classifications of NZE tokens mislabelled as MUSE by classifiers trained on data biased towards MUSE (left) or balanced between MUSE and NZE (right). While the balanced training data resulted in more accurate classifications, the biased model is a much more accurate behavioural model.

The biased model has classifications much more in line with human participants than the balanced one right. The overall accuracy of biased model (F_1=1.00) was also closer to that of human participants (F_1=0.97) than the accuracy of the unbiased model (F_1=0.85).

The model trained on biased data also proved a much better behavioural model for NZE data mislabelled as being MUSE. This can be seen in Figure 6. Again, the biased model is the on the left and the balanced on the right. In particular, the biased model shows the same confusion pattern as human participants, with "had" mislabelled as "head" and "head" mislabelled as "hid". While the model trained on balanced data is more accurate (F_1=0.74, as opposed to F_1=0.48 for the biased model), it is a much closer approximation of participant behaviour; participants were very poor at this task in terms of raw accuracy (F_1=0.56).

Conceptually, we can think of the addition of nodes splitting on dialect as modelling the fact that the more exposure a listener has to a dialect, the more likely they are to make distinctions between that dialect and their own. This is supported by research which shows that listeners are better at identifying dialects the more they have been exposed to them, which has been found in both English (Clopper and Pisoni, 2004; Baker et al., 2009) and Spanish (Díaz-Campos and Navarro-Galisteo, 2009).

4.4 Response time

While greater exposure to multiple dialects may improve comprehension accuracy overall, it does come at a cost. It has been previously found that lexical selection is slower for listeners who have been exposed to multiple dialects (Clopper and Walker, 2016). While all participants in this experiment were exposed to multiple dialects, we can compare their response times for "heed" (which was very similar across both dialects) to those for "had" and "head" tokens (which diverged). A one-way ANOVA of log-transformed response times in milliseconds found a significant effect of word ($F(2,4029)=20.63$, $p < 0.01$). Response times for "heed" ($\mu = 1168.38, \sigma = 541.13$) were the fastest, followed by "had" ($\mu = 1196.62, \sigma = 510.59$) and "head" ($\mu = 1298.16, \sigma = 563.39$). So response times for tokens confusable across dialects was slower than for those tokens not those participants were exposed to less variability for. This can be seen in Figure 7.

This seems even more reasonable given that most conditional inference trees trained on this data, such as the one shown in Figure 8, require travelling through more nodes to classify of "hid", "head" or "head" than they do to classify "heed". Since making a decision at each node takes time, moving through fewer nodes would result in a faster response time.

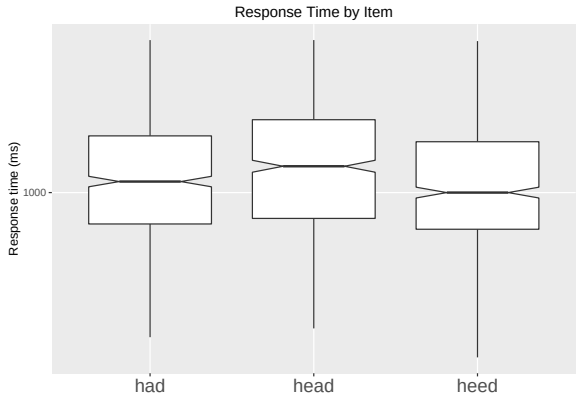

Figure 7: Response time, in milliseconds, by token type. Note that "heed" tokens, which were similar across dialects and thus didn't require dialect-specific learning, were categorized most quickly.

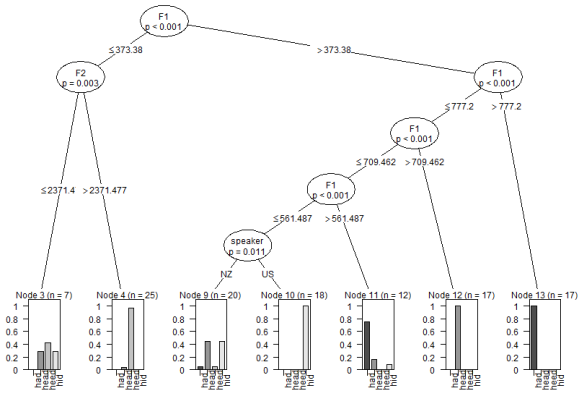

Figure 8: Conditional inference tree trained on balanced MUSE and NZE data. Note that all "heed" classifications occur in nodes 3 and 4 and require travelling through only two other nodes.

5 Discussion

This paper has provided experimental evidence that human listeners maintain a perceptual bias towards their own dialect even after successful perceptual learning. These findings are not surprising given previous findings of social bias towards ones own dialect (Coupland and Bishop, 2007; Floccia et al., 2006), as well as findings in the speech perception literature.

Early work on the perceptual magnet effect (Kuhl, 1991), for example, showed that listeners have a perceptual preference for vowels they are more familiar with, which was attributed to the location of prototypes. Later work on episodic learning casts this preference in terms of a higher density of exemplars in this area (Johnson, 1997; Pierrehumbert, 2001).

Though the modelling results discussed here do fit well with the larger claims of exemplar theory, they are not an explicit extension of it. Note that the balanced conditional inference tree model is not explicitly conditioned on the biased models; each is a separate, independent model. This differs from the Bayesian belief-updating modelling often employed in exemplar theory. Learning can be modelled in this framework in much the same way, however. Training a new model on a dataset that is a superset of the previous training set would allow for "updating" of the model.

The model presented here was weighted to show own-dialect bias. However, there is nothing to stop the weighting from going the other way, which would suggest that it's possible for a listener to be biased *away* from their own dialect. Given that this has previously been observed in phonetics research, this is a feature rather than a bug. With enough exposure to a second dialect, especially if their attitude towards the second dialect is positive, speaker's production and perception can both shift towards the new dialect (Evans and Iverson, 2007; Sumner and Samuel, 2009).

The effect of bias in training data on the classification of conditional inference tree models closely parallels human listener's biases towards their own dialect. While this is desirable from a behavioural modelling perspective, dialectically-biased training data also resulted in lower accuracy on the dialect it was biased against. This suggests that systems trained primarily on one dialect should make the same types of errors on other dialects as a speaker of the training dialect would. In this case, an ASR system trained on MUSE data would make systematic errors on NZE.

This model also provides an alternate way to include speaker dialect during automatic speech recognition. Rather than training separate models on each dialect (Telaar and Fuhs, 2013; Najafian et al., 2014), dialect can be included within decision-tree based acoustic models (Young et al., 1994). This may be especially helpful for closely related dialects or varieties that share phones. It may, however, come with a cost. Given the response time data discussed in Section 4.4, an accurate behavioural model capable of multi-dialect recognition would respond more slowly to tokens that are confusable between the dialects it was trained on.

References

Wendy Baker, David Eddington, and Lyndsey Nay. 2009. Dialect identification: The effects of region of origin and amount of experience. *American Speech*, 84(1):48–71.

Cynthia G. Clopper and David B. Pisoni. 2004. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change*, 16(01):31–48.

Cynthia G. Clopper and Abby Walker. 2016. Effects of lexical competition and dialect exposure on phonological priming. *Language and Speech*, page 0023830916643737.

Nikolas Coupland and Hywel Bishop. 2007. Ideologised values for British accents. *Journal of sociolinguistics*, 11(1):74–93.

Manuel Díaz-Campos and Inmaculada Navarro-Galisteo. 2009. Perceptual categorization of dialect variation in spanish. In *Selected Proceedings of the 11th Hispanic Linguistics Symposium*, pages 179–195.

Bronwen G. Evans and Paul Iverson. 2007. Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, 121(6):3814–3826.

Caroline Floccia, Jeremy Goslin, Frédérique Girard, and Gabrielle Konopczynski. 2006. Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, 32(5):1276.

Robert Allen Fox. 1974. An experiment in cross-dialect vowel perception. In *Tenth Regional Meeting–Chicago Linguistic Society*, pages 178–185.

Jennifer Hay and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics*, 48(4):865–892.

Jen Hay, Katie Drager, and Paul Warren. 2010. Short-term exposure to one dialect affects processing of another. *Language and speech*, 53(4):447–471.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Keith Johnson, Elizabeth A. Strand, and Mariapaola D'Imperio. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27(4):359–384.

Keith Johnson. 1997. Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing*, pages 145–165.

Patricia K. Kuhl. 1991. Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Attention, Perception, & Psychophysics*, 50(2):93–107.

William Labov, Sharon Ash, and Charles Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Christian Langstrof. 2009. On the role of vowel duration in the New Zealand English front vowel shift. *Language Variation and Change*, 21(3):437.

Samantha Lyle. 2008. *Dialect variation in stop consonant voicing*. Ph.D. thesis, The Ohio State University.

Maryam Najafian, Saeid Safavi, Abualsoud Hanani, and Martin Russell. 2014. Acoustic model selection using limited data for accent robust speech recognition. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 1786–1790. IEEE.

Nancy Niedzielski. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, 18(1):62–85.

Janet Pierrehumbert. 2001. lenition and contrast. *Frequency and the emergence of linguistic structure*, 45:137.

Gijsbert Stoet. 2010. Psytoolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4):1096–1104.

Meghan Sumner and Arthur G. Samuel. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4):487–501.

Meghan Sumner. 2011. The role of variation in the perception of accented speech. *Cognition*, 119(1):131–136.

Dominic Telaar and Mark C. Fuhs. 2013. Accent- and speaker-specific polyphone decision trees for non-native speech recognition. In *INTERSPEECH*, pages 3313–3316.

Catherine Watson. 2014. Mappings between vocal tract area functions, vocal tract resonances and speech formants for multiple speakers. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Richard Wright. 2004. A review of perceptual cues and cue robustness. *Phonetically based phonology*, pages 34–57.

Steve J. Young, Julian J. Odell, and Philip C. Woodland. 1994. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics.