Speaker Dialect is a Necessary Feature to Model Perceptual Accent Adaptation in Humans

Rachael Tatman
University of Washington
Seattle, WA 98105
rctatman@uw.edu

Abstract

In accent adaptation–adjusting existing ASR to recognize novel accents–systems commonly make use of dialect labels. This project models parallel experimental behavioral data, where human listeners were trained to categorize speech sounds from a novel dialect. Explicitly including dialect information in the model allowed the classifier to better simulate the behavioral results.

1 Introduction

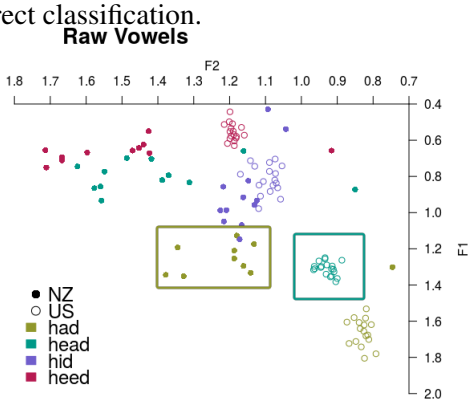
Feature selection during ASR is often automatic and based on the acoustic waveform (Kumar and Andreou, 1998; Ni et al., 2015). In accent adaptation, however, ASR researchers have successfully used accent information during model training (Humphries et al., 1996; Najafian et al., 2014). For this project, a model was constructed of the behavior of human participants completing a classification tasks analogous to automatic accent adaptation. This project investigated whether including social information (in this case dialect region) about speakers would result in more human-like classification.

2 Data

Acoustic data was taken from New Zealand English (NZE) (Watson, 2014) and American English (AmE). These dialects were chosen due to the existence of confusable mergers–NZE "had" and AmE "head" are homophones (Hay et al., 2008). The Neary normalized (Nearey, 1978) acoustic data is shown in Figure 1.

American participants were trained on NZE vowels and then asked to classify a second set of NZE vowels. However, only half the participants were aware that they were listening to NZE.

Figure 1: True classification of vowels in test set, by dialect and context. Box colors correspond to correct classification.



The other half were told that they were listening to AmE vowels. This was done in order to determine the role of dialect information on human classification of speech sounds. The results were striking–participants relied heavily on the given dialect information (even when it was incorrect) in determining how to classify speech sounds (Tatman, 2016). This suggests that human-like automatic classification of speech sounds should also make use of dialect information.

3 Models

The models used here are conditional inference trees, implemented in R using the package partykit (Hothorn and Zeileis, 2014). While not standard for segment classification in ASR, conditional inference trees have the benefit of being easily interpretable, which is desirable for behavioural models. Both classifiers were trained in the same way, with word as the classification output and Neary-normalized F1 and F2 and, in the second model, speaker dialect as features. Models were trained and evaluated using sub-sampling cross-validation, with half of the data randomly selected

Figure 2: Classification on data with correct dialect info, both models.

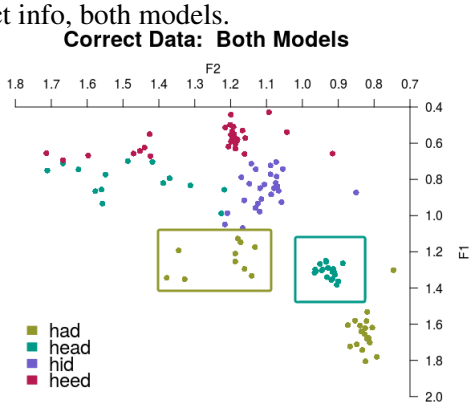
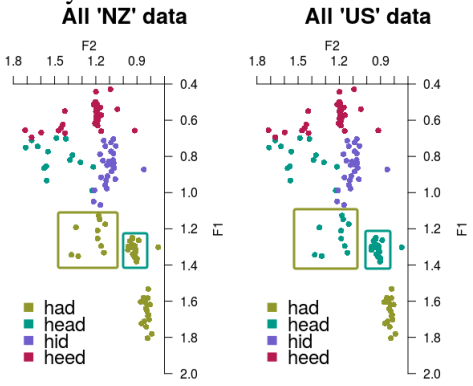


Figure 3: Classification of model with dialect, on incorrectly labelled data.



for training and the rest used for validation. Both models achieved the same classification results on the original data with the correct dialect information (Figure 2), with a balanced accuracy of 0.86. The difference between models only becomes apparent when the dialect of a token is labeled incorrectly. The first model, which includes no dialect features, does not change its classifications. However, the second model correctly mirrors human behaviour–it classifies tokens as if they were from the labeled dialect, as can be seen in Figure 3. Note that the classification of NZE "had" and AmE "head" tokens depends on the labeled dialect.

4 Discussion

Behavioural experiments are not generally part of feature selection. But this project suggests that this may be appropriate when it's desirable to model human behaviour. If model selection between the two classifiers described above rested only on 1) acoustic information and 2) parsimony, then the first model would have been the better choice. Unfortunately, this would have missed capturing an important fact about human accent adaptation, which was captured by the inclusion of speaker dialect as a feature.

This project suggests that explicitly including a speaker's dialect information during ASR may help to provide more human-like recognition, especially where there is possible confusion between dialects. This follows with other work which has shown that accent-specific models can improve the accuracy of ASR (Najafian et al., 2014).

References

Jennifer Hay, Margaret Maclagan, and Elizabeth Gordon. 2008. *New Zealand English*. Edinburgh University Press.

Torsten Hothorn and Achim Zeileis. 2014. partykit: A modular toolkit for recursive partytioning in r. Technical report, Working Papers in Economics and Statistics.

Jason Humphries, Philip Woodland, and D Pearce. 1996. Using accent-specific pronunciation modelling for robust speech recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2324–2327. IEEE.

Nagendra Kumar and Andreas Andreou. 1998. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26(4):283–297.

Maryam Najafian, Andrea DeMarco, Stephen Cox, and Martin Russell. 2014. Unsupervised model selection for recognition of regional accented speech. In *INTERSPEECH*, pages 2967–2971.

Terrance Nearey. 1978. *Phonetic feature systems for vowels*, volume 77. Indiana University Linguistics Club.

Chongjia Ni, Lei Wang, Haibo Liu, Cheung-Chi Leung, Li Lu, and Bin Ma. 2015. Submodular data selection with acoustic and phonetic features for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4629–4633. IEEE.

Rachael Tatman. 2016. Listening with american ears: Using social information in perceptual learning. In *Experimental Approaches to Perception and Production of Language Variation*. University of Vienna.

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*.