# Classifying and Clustering Dialects of North American English

**Keelan Evanini** KEELAN2@LING.UPENN.EDU
Department of Linguistics, Unversity of Pennsylvania, Philadelphia, PA 19104 USA

## Abstract

This paper presents the results of experiments in which machine learning techniques were applied to the problem of determining regional dialect boundaries. Specifically, decision trees classification and k-means clustering were applied to a corpus of phonetic measurements taken from a large survey of North American English vowels. Pairwise classification and clustering experiments were done for all combinations of ten dialect regions determined by dialectologists. The results show which of these dialect regions are most distinct and similar, suggesting which of the distinctions that are usually used by linguists are the most meaningful. Furthermore, the classification trees are analyzed to show which vowel formants are most informative for each dialect region.

## 1. Introduction

One of the most difficult theoretical questions for dialectologists is determining a principled way to partition the linguistic area into dialect regions. Traditional work in dialectology has relied on subjectively determined bundles of isoglosses that are often influenced by existing notions of the boundaries, both for lexical, e.g. (Carver, 1987), and phonological, e.g. (Kurath & McDavid, 1961), isoglosses. The problem is especially difficult since there is always some amount of overlap (both in lexicon and phonology) between neighboring regions. This work stems from the assumption that the application of machine learning techniques to the problem can help dialectologists make better informed, pre-theoretic decisions about dialect boundaries.

Some recent approaches on dialect clustering have used techniques such as mutual information (Nagy et al., 2005) and Levenshtein distances (Nerbonne & Heeringa, 1997). However, all such approaches have worked with data that has been abstracted away from the original source, i.e. transcriptions or annotations from linguistic atlases. The present work is novel in that it uses acoustic phonetic measurements taken from field recordings. Furthermore, it is the first study to consider all of the dialects of North American English as a whole.

Presented at *North East Student Colloquium on Artificial Intelligence (NESCAI)*, 2008. 

## 2. ANAE Corpus

The data source for this project is the *Atlas of North American English* (Labov et al., 2006), henceforth ANAE. ANAE is by far the most comprehensive study of regional dialect variation in North America. It contains at least 300 vowel formant measurements (F1 and F2 normalized using a log-mean procedure (Nearey, 1977)) for each of 439 speakers representing all dialect regions (at least two speakers were sampled randomly from every Mean Statistical Area in North America with at least 50,000 residents). Table 1 provides a breakdown of the ANAE speakers by dialect region (as determined manually by the ANAE authors by taking into account the sound changes in progress in each region and the homogeneity and consistency of the isoglosses).

*Table 1.* Dialect regions in ANAE

| Dialect Region | Abbreviation | # Speakers |
|---|---|---|
| North | N | 126 |
| South | S | 83 |
| Midland | M | 65 |
| West | W | 51 |
| Canada | CA | 33 |
| Transitional | T | 26 |
| Western PA | WPA | 15 |
| Mid-Atlantic | MA | 15 |
| Eastern New England | ENE | 11 |
| Southeast | SE | 10 |
| New York City | NYC | 5 |
| Total | | 439 |

For the experiments in Section 3, each speaker is

treated as an instance; the Transitional speakers were excluded since they do not form a contiguous region. The feature vector for each speaker consists of F1 and F2 means for all 23 vowels in American English, as defined by the ANAE authors. The tasks thus involve 10 classes (dialect regions), 413 instances (speakers) and 46 features (vowel formant means).

## 3. Experiments and Results

Two different machine learning approaches were taken to determine how similar / distinct the 10 dialect regions defined by ANAE are: a supervised classification task using decision trees (with 10-fold cross validation) and an unsupervised clustering task using k-means (averaged results of 50 trials). For each approach, binary classification / clustering (where $k = 2$) tasks were done for each of the 45 dialect pairs. Furthermore, binary classification was done for each of the regions against all other speakers not from that region.

Table 4 in Appendix A presents the overall classification results for each region. Two scoring methods (average F-measure and $\chi^2$ value of the 2x2 confusion matrix) provide the same ranking. The results coincide well with what is already known from work in dialectology: the three regions that the classifier recognizes as most distinct from all the others are all undergoing vowel changes that are unique to those regions (e.g. the fronting of /o/ in the Northern Cities Shift does not take place elsewhere, nor does the backing of /æ/ in the Canadian Shift—see (Labov et al., 2006) for descriptions of all of the specific sound changes discussed in this paper). On the other hand, the dialects that are least easily separated from the rest are undergoing some sound changes that are also taking place in other regions, e.g. the strongest fronting of /ow/ is found in WPA and MA, and the merger of /o/ and /oh/ is a salient characteristic of M, W, and WPA. Table 5 in Appendix B lists the most informative vowel formant for each region based on the first decision tree split that was made for all of the classification tasks in Table 4.

Tables 2 and 3 present the confusion matrices for three illustrative examples of the pairwise classification task. The first task, N vs. S, is commonly regarded as the easiest, since speakers in the two regions are undergoing large-scale vowel shifts in opposite directions (the Northern Cities Shift and the Southern Shift, respectively). Both tasks were able to separate the speakers from the two regions quite well, with results comparable to the 8% error rate attained by (Miller & Trischitta, 1996) using a linear discriminant on mean cepstral and duration features. The second comparison presents the type of unexpected positive result that these tasks were undertaken to show: CA and MA are not usually considered together by dialectologists, but their decision tree classifier had near perfect performance (0.95 average F-measure). The decision tree for CA vs. MA has only one split: if the 2nd formant of /aw/ is greater than 1701 Hz classify as MA, otherwise classify as CA. This makes good linguistic sense, because /aw/ is moving in opposite directions in the two regions: the Midland has the strongest fronting of /aw/ of any region, and it often becomes the diphthong [eɔ], whereas one of the most salient features of Canadian English is Canadian Raising, in which the nucleus of /aw/ is raised and backed. Finally, the third pairwise contrast in Tables 2 and 3 shows a case in which the two regions overlap substantially. Both the pairwise classification and clustering tasks between M and W had poor performance, indicating that their vowel systems are quite similar. Instances of poor performance like this provide examples of cases in which the traditional dialect divisions should be reconsidered.

*Table 2.* Confusion matrices for three classification tasks

| true → | N | S | | CA | MA | | M | W |
|---|---|---|---|---|---|---|---|---|
| N | 116 | 10 | CA | 32 | 1 | M | 42 | 23 |
| S | 9 | 74 | MA | 1 | 14 | W | 22 | 29 |

*Table 3.* Confusion matricies for three clustering tasks

| true → | N | S | | CA | MA | | M | W |
|---|---|---|---|---|---|---|---|---|
| N | 119 | 2 | CA | 26 | 1 | M | 38 | 17 |
| S | 7 | 81 | MA | 7 | 14 | W | 27 | 34 |

Table 6 in Appendix C presents the results for all pairwise classification and clustering tasks; the results from the two methods overlap to a large extent. Dialectologists should consider combining pairs of regions such as SE and M that are determined by both methods to have a large amount of overlap.

## 4. Further Work

Experiments are currently underway to increase $k$ in the k-means clustering tasks to determine which number of regions results in the highest homogeneity in each region. Furthermore, hierarchical clustering experiments will help inform dialectologists how best to label a given speaker. Finally, it is recognized that limiting the feature set to only F1 and F2 measurements does not provide a comprehensive view of dialect variation. Work is underway to transcribe the ANAE interviews so that forced alignment can be used to also extract other useful phonetic measurements, such as F3, duration, and F0.

## References


Carver, C. M. (1987). *American regional dialects: A word geography.* Ann Arbor: University of Michigan Press.

Kurath, H., & McDavid, R. I. (1961). *The pronunciation of English in the Atlantic states.* Ann Arbor: University of Michigan Press.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English.* Mouton de Gruyter.

Miller, D. R., & Trischitta, J. (1996). Statistical dialect classification based on mean phonetic features. *Proc. ICSLP*, 2025–2027.

Nagy, N., Zhang, X., Nagy, G., & Schneider, E. (2005). A quantitative categorization of phonemic dialect features in context. In A. Dey (Ed.), *CONTEXT 2005 lecture notes in artificial intelligence 3554*, 326–338. Berlin Heidelberg: Springer-Verlag.

Nearey, T. (1977). *Phonetic feature system for vowels.* Doctoral dissertation, University of Connecticut.

Nerbonne, J., & Heeringa, W. (1997). Measuring dialect distance phonetically. *Proceedings of the third meeting of the ACL special interest group in computational phonology* (pp. 11–18).

## A. Overall classification results

*Table 4.* Classification results for each region vs. all others

| Region | Avg. F-measure | $\chi^2$ |
|---|---|---|
| N | .82 | 174.0 |
| S | .81 | 162.2 |
| CA | .80 | 156.0 |
| SE | .68 | 55.3 |
| NYC | .68 | 52.7 |
| M | .67 | 49.3 |
| ENE | .65 | 38.3 |
| MA | .64 | 35.2 |
| W | .64 | 31.9 |
| WPA | .52 | 0.51 |

## B. Most informative formants for each region

*Table 5.* Features and values of the first splits made by the decision tree classifier for each region vs. all other regions

| Region | Formant split |
|---|---|
| CA | OWC1 <= 581 |
| ENE | UWF2 <= 1587 |
| MA | OH1 <= 665 |
| M | OWC2 > 1180 |
| N | O2 > 1395 |
| NYC | OH1 <= 657 |
| SE | OHR2 <= 881 |
| S | EYC1 > 617 |
| W | O2 < 1396 |
| WPA | OWR1 <= 469 |

## C. All pairwise classification and clustering resuls

*Table 6.* $\chi^2$ values for pairwise decision trees classification results (upper right half) and k-means clustering results (lower left half); top ten $\chi^2$ values for each task in **bold**, values not significant at $\alpha = 0.05$ in *italics*

| | N | S | M | W | CA | WPA | MA | ENE | SE | NYC |
|---|---|---|---|---|---|---|---|---|---|---|
| N | – | **137.3** | **66.6** | **89.0** | **107.8** | 59.4 | 47.1 | **75.7** | **68.0** | 33.0 |
| S | **173.7** | – | 48.7 | **71.4** | **71.2** | 11.0 | 17.8 | 57.1 | 29.2 | **69.6** |
| M | **78.5** | 9.9 | – | 5.3 | 51.2 | 12.8 | 26.0 | **60.3** | *0.1* | 40.7 |
| W | **48.4** | **100.6** | 6.9 | – | 57.2 | 20.8 | 27.2 | 36.6 | 14.3 | 43.9 |
| CA | **29.4** | **90.4** | **28.6** | *2.1* | – | 14.0 | 39.1 | 7.3 | 25.0 | *2.5* |
| WPA | 13.1 | 12.6 | *0.4* | *0.7* | 14.3 | – | 8.6 | 9.5 | *2.8* | 15.0 |
| MA | 17.7 | 13.5 | *0.6* | 4.2 | 22.0 | *1.9* | – | 9.5 | 18.1 | *0.1* |
| ENE | 4.3 | **85.1** | **26.3** | **29.0** | 17.2 | 22.1 | 22.2 | – | 10.8 | *0.3* |
| SE | 10.8 | *3.1* | *2.7* | 10.8 | 13.9 | 4.4 | *3.4* | 19.1 | – | 5.0 |
| NYC | *0.3* | 4.0 | 5.0 | *0.3* | 6.0 | 7.9 | 6.7 | *2.5* | *3.4* | – |