

A Focused Analysis of Interpretable Phonetic Dialect Classification¹

Henry Nomeland[†]

Abstract. While analyzing dialect variation has historically required careful transcription and specialist analysis, variations can now be identified using machine learning. This research trains classification models such as random forests on features such as vowel formants to classify speakers by their locations of origin. Interpretation methods which account for intercorrelated features then identify which phonetic characteristics the models most heavily rely on. These methods identify the frequent /O/ and /A/ distinction between Americans and Britons in words like call, the short-a raising common among New Yorkers, and several other characteristics, demonstrating a novel and objective method for understanding regional speech.

1. Introduction. This paper represents a condensed edition of an undergraduate thesis, *Interpretable Dialect Classification: A Statistical Modeling Pipeline for Identifying Phonetic Dialect Features*. Sections which were not a focus of the corresponding presentation at the Workshop in General Linguistics 18 have been omitted. Similarly less relevant sections have been abbreviated. These omissions reflect an emphasis on the linguistic components of the research while minimizing discussion on modeling and statistical considerations. Details regarding collinearity, historical context, cross validation, and model interpretation have been omitted or reduced.

The most recent and novel approaches in the study of linguistic variation have begun to incorporate speech processing technologies and machine learning methods (Wieling & Nerbonne 2015; Jeszenszky et al. 2021). This paper deals with the systematic identification of phonetic differences between English-speaking individuals from different regions. Specifically, statistical classification models will be utilized to distinguish dialects from one another, and identify features of speech that differ between speakers at the acoustic level. A large segment of linguistics literature has focused on the many different ways that sounds are produced to explain acoustic variation. Some of these studies analyze very particular phonemes and particular places, such as the presence or absence of a vowel merger within a particular population (Thiel & Dinkin 2020; Baranowski 2022). Others consider sets of acoustic features, such as several vowel sounds, that have undergone changes in related ways (Voeten 2021; D’Onofrio & Benheim 2020). What many of these studies take for granted is the salience of features in comparison to one another.

Whether one characteristic is more or less accentuated or common among a population is often an afterthought of otherwise careful analysis of individual speakers and groups. Likewise, the importance of one speech characteristic over another in terms of defining a dialect is often a subjective debate between specialists, especially as dialects constantly shift, converge, and influence one another (Eide & Afarli 2020; Ross et al. 2021).

Few studies have sought more systematic, objective and data-driven approaches for identifying phonetic characteristics of English dialects. For the demonstration cases provided in this paper, a small subset of vowel sounds were identified and measured from short audio

¹ This work is a shorter version of an undergraduate thesis titled *Interpretable Dialect Classification: A Statistical Modeling Pipeline for Identifying Phonetic Dialect Features*, which was completed in the Statistics Department at the University of Wisconsin-Madison in the spring of 2024.

[†] UW-Madison Statistics Department; advisor Garvesh Raskutti

samples. Vowel sounds are extremely important in the perception of dialectal variation, with most undergoing significant changes in quality over time (Tanner et al. 2022). Their sonorant nature makes them easy to measure as frequencies, and the scalar nature of vowels lends them to greater amounts of variance between speakers when compared to consonants (Ladefoged 1993). These vowel sounds were used to train models that classify English speakers by their locations of origin. Different datasets define this location differently from one another, but generally this location label is the place that the individual in question has been living in for the longest period of time.

After training, models will be interpreted in an attempt to identify which vowel sounds and which characteristics of those vowel sounds prove to be the most influential in the classification process. This objective of interpretation is the primary goal of interpretable machine learning (ML), a relatively novel paradigm that seeks to balance high performance models with models that are easy to deconstruct and understand. While complex models often have strong predictive capacities, they are often ‘black box’ models, providing few relevant domain insights into how particular predictions and classifications are actually made (Murdoch et al. 2019). In order to align with the goals of interpretability, some design decisions may forego potential improvements in model accuracy in exchange for improvements in model interpretability.

Regions and dialects are not made up of homogenous groups of speakers all pronouncing every phoneme with exacting precision. Everyone has a unique voice as a result of their upbringing, culture, and physiology, and every acoustic feature present in a region won't be found in every speaker living in that region. In this respect, one can say that everyone has their own unique and personalized idiolect (Ke 2004). For this reason, it must be recognized that certain features are more important than others when identifying where someone is from. Some features may be very common and distinct among a population. Others might be as common but more subtle, while others still may be relatively rare but extremely prominent and unique, making classification easy for a small subset of a region's population. This variation makes identifying and comparing the importance of each feature a unique challenge, and interpretable machine learning a suitable solution. It also means that the predictions made by the classification models proposed henceforth are not predictions of an individual's dialect. Some dialects exist across regional boundaries, some regions contain many very disparate dialects, and some dialects are subsets of broader dialect families. In these ways, dialects can be said to overlap in many ways, making their categorization a unique challenge (Chittaragi et al. 2017).

Unlike many other studies which have used ML methods in the field of dialectology, this paper does not seek to identify dialect groups from a population of speakers. It also does not intend to identify all relevant features of any particular dialect. Rather, the intention is to better understand how people and machines predict where individuals are from based on their speech. Given a particular speaker and a general idea of the places they might be from, what features of speech are the biggest giveaways in terms of whether they are from one location or another?

2. Literature review. In Themistocleous (2019) several models, some more interpretable than others, were implemented in order to distinguish Athenian Greek speakers from Cypriot Greek speakers. While relatively complex models such as deep neural networks and random forests were used for training, the input variables were very simple, consisting of a small number of experimentally collected sonorant consonant sounds from which the author extracted formant measurements. Trained on this very small subset of acoustic measurements, the best

performing model classified speakers with an accuracy well over 65%. In discussion the author notes that having a highly controlled experimental framework reduces noise, resulting in model performances which are comparable to pipelines based on larger amounts of more variable features. They also note that the selection of specific features for analysis can allow for a better understanding of how humans perceive speech variations compared to more abstract features.

The benefits of selecting interpretable features and models are further explored by Strycharczuk et al. (2020) who classify various English speakers native to Northern England by the particular city that they are from. The authors utilize random forest interpretation methods and a subset of vowel formants collected uniformly across all speakers. They go one step further than most other dialect classification studies by calculating variable importance metrics for each independent variable followed by a comparison of the variables with the highest importance for each classification against professional qualitative assessments of the audio samples. While this analysis demonstrates the potential of variable importance calculations in dialectology research, the authors report a very small number of important features, and further validation could be done to confirm that important features align with current understandings of dialect variation.

Of the existing studies that have calculated variable importance metrics following dialect classification, very few have rigorously compared the feature rankings with domain knowledge to validate the accuracy the interpretations. This paper places higher relative importance on model interpretability than previous works, and cross-validates the results of model interpretations with existing linguistic knowledge of regional speech variation. Many existing state-of-the-art dialect classification models make use of fine-grain acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) to achieve high classification performances even in cases with many different dialects to consider (Humayun et al. 2024). While other studies have utilized less abstract features to classify dialects, the goals of such studies are typically to demonstrate that such features are correlated with a geographic grouping, and could potentially improve more complicated classifiers (Lou & Ren 2021; Themistocleous 2019). While some have calculated variable importance metrics of model features, this is normally an afterthought of model development and performance.

3. Figures, drawings, and tables.

3.1. DATASET CONSIDERATION. In the discipline of speaker recognition, evaluations are typically divided into text-dependent and text-independent cases. Speaker recognition can serve as a good point of comparison to dialect classification as both seek to classify speech samples using acoustic information. Text-dependent recognition requires the speaker to utter a set sequence of words that the model has been trained to interpret. This contrasts with text-independent recognition, where speakers provide the model with arbitrary utterances. In the text-dependent case the degree of variance from sample to sample is much smaller, giving researchers far more control. This also allows the model to be trained on much less data, including shorter utterances with fewer words (Hebert 2008). Studies that perform dialect classification on text-independent datasets deal with much more noise in the input, necessitating the need for larger datasets and more complicated architectures (Najafian et al. 2016).

With interpretability in mind, a model trained using minimal features representing particular aspects of particular phonemes is preferred over an arbitrarily fine-grained feature set. This makes a text-dependent method ideal, as it allows for less complex and more targeted

models. Particular phonemes can be pronounced very differently from word to word. For example, the aspirate consonant is highly dependent on the succeeding vowel, and many stops can be pronounced with or without aspiration. Vowels can change from context to context depending on stress, and fricatives can have varying degrees of voicelessness depending on their location within a word (Ladefoged 1993). If these allophones are treated as the same sound, the derived features will have far greater variance across speakers than if they were calculated from sounds derived from the same word and sentence for every speaker and every sample. Even if samples were transcribed more broadly to prevent this issue, the amount of data needed would be far greater than in a text-dependent method which can isolate for a small number of highly consistent phonemes all spoken in the same words and contexts. This makes the text-dependent method uniquely useful for interpretability. While a text-independent method allows for more flexibility in gathering data, the far less complex nature of text-dependent modeling makes large datasets less of a requirement.

With the goal of text-dependent modeling in mind, demonstration datasets must consist of many speakers saying the same utterance, ideally under the same or similar conditions. As the goal is region classification, the speakers must be labeled with a distinct geographic tag. Voice samples have unique limitations due to their sensitive nature. Speech can be used as an identifier of study participants, and there is no consensus on whether such data is de-identifiable (Cychosz 2020). As a result, the scale of high-quality datasets of speech utterances is rather exiguous when compared to other scientific domains. Luckily, some efforts have been made to make certain studies and corpuses available.

3.2. DEMONSTRATION DATASETS. Two datasets which contained geographically labeled, text-dependent audio samples of speakers were identified for analysis. The first is the George Mason University (GMU) Speech Accent Archive, a database dedicated to collecting audio samples from an extremely wide variety of speakers (George Mason University 2010). Audio samples from this dataset consist of a 69-word utterance. Demographics such as gender and age vary widely but consistently across regions. There are a total of 431 American and 73 British speakers present in the dataset. A classification of British speakers against American speakers was chosen due to the high amounts of data from these areas.

The second dataset is the TIMIT acoustic-phonetic continuous speech corpus, a database of American English speakers from several regions of the United States (Garofolo et al. 1993). Samples consist of utterances recorded across two tracks which necessitated additional preprocessing. There are a total of 630 American speakers in the dataset of which 46 are from New York City. A classification of New York City speakers against other American speakers was chosen due to the relatively unique nature of New York City speech relative to the other regions. Classification performances reported in Appendix A show that speakers from the other regions are much more difficult to identify.

3.3. FEATURE SELECTION. It is well known in the machine learning community that models trained on smaller numbers of controlled variables are generally easier to interpret. This necessarily trades off with model accuracy (Ghose & Ravindran 2020). In order to minimize the damage done by selecting a small feature set, it is best to choose a set of features which contain a large amount of descriptive information. As previously explained, vowels are often highly variable across linguistic regions. The average F1 and F2 formant values over the course of a vowel can succinctly summarize the resonant frequencies produced by the specific vocal tract shapes used to make those vowels (Ladefoged 1993). This results in two simple values measured in hertz which define the sounds to be analyzed. Generally speaking the F1 frequency

is associated with how high or low the tongue is positioned in the mouth during production, while the F2 frequency is associated with how far forward or back the tongue is (Ladefoged 1993). Higher-pitched formant frequencies can encode some information about vowel quality but the majority is encoded in the two lowest values. For the sake of simplicity only monophthongs occurring as the first vowel in a word were included as features.

To extract formant values the intervals of each vowel sound needed to be derived. To do this, the Montreal Forced Aligner was employed to align the dataset transcripts automatically with each individual audio sample (McAuliffe et al. 2017). From here, a Praat implementation of a short-term spectral analysis was used to find high points in the spectrum and derive formant tracks (Boersma & Weenink 2017). This Praat functionality was implemented in Python using the Parselmouth package (Jadoul et al. 2018). Midpoints were extracted from the F1 and F2 tracks to get two individual frequency measurements associated with each relevant vowel in the data. These are the measurements used as inputs to the classification models.

3.4. MODEL SELECTION. After some preliminary research into models that have been used for dialect classification in the past, and excluding neural network and Markov chain models for the sake of interpretability, four primary models were considered and implemented in Python using scikit-learn (Pedregosa et al. 2011). These included logistic regression, random forests, simple decision trees, and k-nearest neighbors. Preliminary results regarding support vector machines and other algorithms were promising and may be analyzed in future work. Performances for the four primary models were calculated using leave-one-out cross-validation. Hyper-parameters for each model were optimized using scikit-learn search algorithms.

Feature scaling was performed due to features being on a wide range of different scales. As a result of overwhelmingly favorable calculations of model accuracy as reported in Figure 4, min-max normalization was chosen over z-score standardization for all modeling.

Of the models evaluated, the random forest and regularized logistic regression models were the best performing. Logistic regression models tend to be more prone to overfitting when modeling high-dimensional data. This issue can be partially addressed by using regularization (James et al. 2023). Analyzing regularized logistic regression alone is the simplest solution to exploring these demonstration datasets, but because this paper seeks to propose a pipeline which allows for the inclusion of new datasets and features, random forests will also be considered. Previous studies have concluded that random forests and their ability to deal with complex interactions make them uniquely suited for classifying datasets with more complicated feature sets (Strycharczuk et al. 2020; Humayun et al. 2024). It is additionally useful to have random forests as an option as they can easily work with categorical variables such as vocalic quality, aspiration, or fronting (Schaeffler 2005).

3.5. MODEL TRAINING AND CROSS VALIDATION. The primary objective of all of the models implemented using this pipeline is to differentiate individuals from a particular location from a larger population of speakers. Due to the nature of this binary classification, the group of interest will almost always have a much smaller sample size compared to the remainder of the population. Machine learning algorithms such as decision trees are designed to maximize accuracy regardless of underlying population distributions, and they are likely to be biased towards the majority class as a high accuracy can be achieved by simply assigning most or all samples a majority class label (He & Ma 2013). To deal with this issue it is a common practice to balance class sizes through undersampling, oversampling, or a combination of the two

practices. In this particular use case where datasets are relatively small, it is important to maximize the use of available data for training.

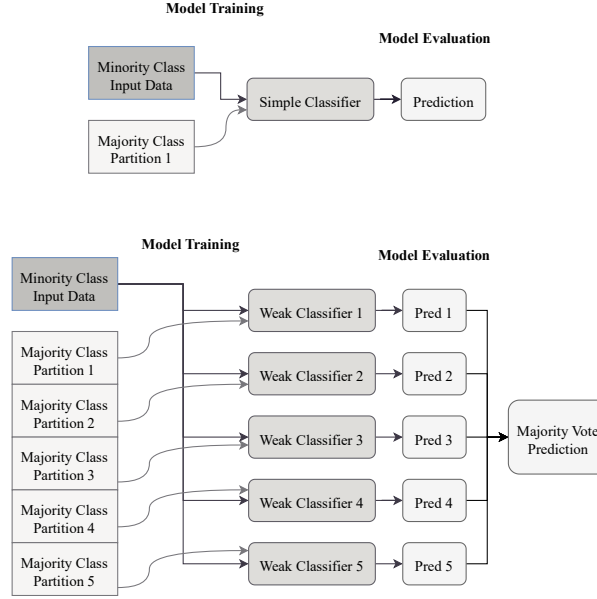


Figure 1. A demonstration of majority class bias in logistic regression.

One particularly effective way of achieving this is by randomly undersampling the majority class several times and training several models to be combined into a single ensemble model (Khan et al. 2024). While complex boosting algorithms such as EasyEnsemble and RUSBoost have the potential for better performance by learning from individual training examples multiple times, other methods can achieve similar goals with relative simplicity. This paper uses the method of partitioning, which involves randomly drawing subsets from the majority class without replacement as seen in Figure 1. Weak learners are then trained from this data paired with the minority class samples, and predictions take the class which is predicted by the most weak learners (Khan et al. 2024). This algorithm allows for all or nearly all samples to be included in the ensemble in some capacity, and its simplicity allows for greater control regarding cross validation and variable importance techniques.

Many classification pipelines split training and testing data prior to modeling. This omits a large portion of dataset samples that could be used for training and may result in sub-optimal performance measurements as a result of testing on a small fraction of available data. When working with small datasets such as these it is better to utilize leaveone-out cross-validation, which trains n models each trained on $n-1$ samples which are each tested on the held-out sample (James et al. 2023). In the case of classification the results of each of these individual predictions are then aggregated into a final performance metric. This minimizes bias and reduces randomness inherent in deciding a train-test split.

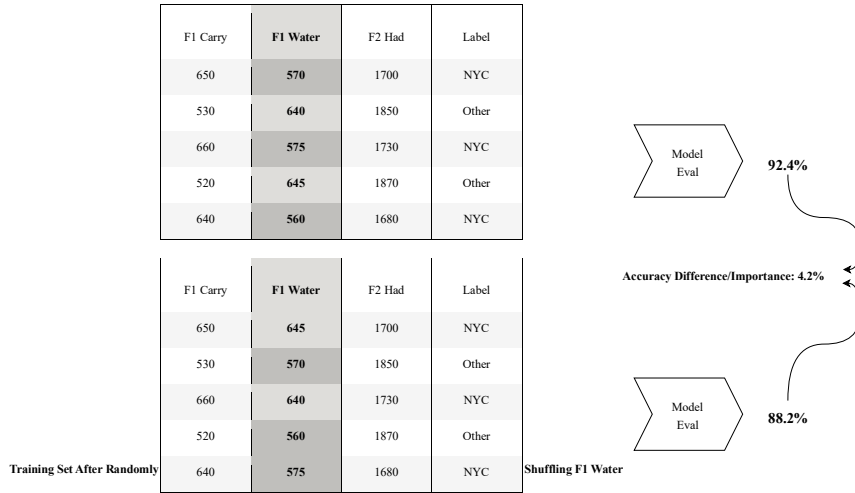


Figure 2. A simple demonstrative example of a permutation importance calculation

3.6. INTERPRETATION METHODS. Both regression and random forest models have model-specific interpretation methods that are routinely used for identifying important independent variables. For regression, the coefficients of each independent variable after training approximately correspond to that variable's relationship to the model's output. For random forests, importance metrics include mean decrease impurity (MDI) and permutation importance. Impurity importance metrics are derived from the structure of tree-based models, with each feature score based on the contribution of that feature to sorting out the input samples (Nembrini et al. 2018). The permutation method looks at the baseline accuracy of a model trained on all available features and compares it to an inferior model where a single feature has been randomly permuted in place (see Figure 2 above). The importance of the scrambled feature is calculated by taking the difference in accuracy between the strong base model and the permuted weak model (Wei et al. 2015).

For all three importance metrics, independent importance scores are calculated for each partition of the majority class using the corresponding weak classifier. The importance metrics are then averaged across the many different weak classifier calculations to provide a single aggregate ranking of all of the model features as depicted in Figure 3. These methods ensure that the importance of each feature is calculated using the maximum amount of data.

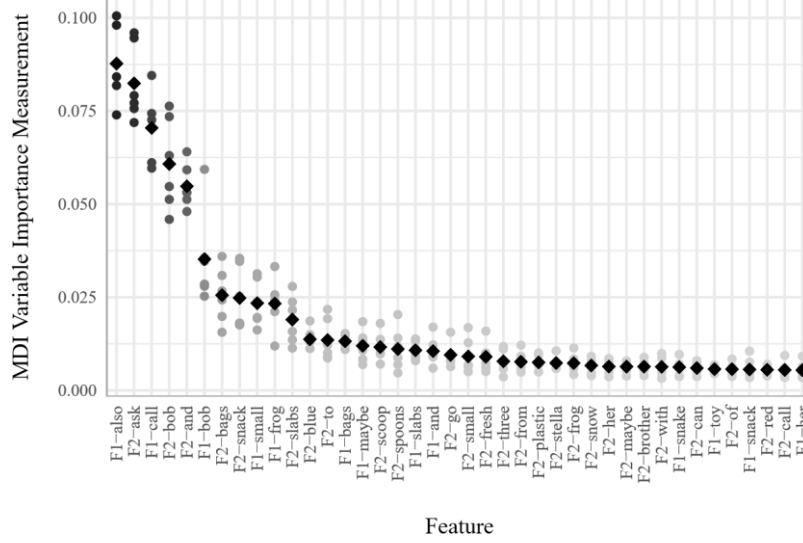


Figure 3. The features of the UK random forest classification ranked by MDI score. Diamonds represent the average MDI across all weak classifiers.

4. Results and analysis.

4.1. MODELING RESULTS. The earlier section on dataset selection provided several reasons why different locations can have very similar sounding speakers, and why speakers from the same location can sound extremely different from one another. In addition, this paper focuses on interpretability as opposed to maximizing model performance. For these reasons, perfect regional classification of individuals by their voices is not a viable goal.

Despite these limitations, the best performing models prepared for this paper were able to distinguish United Kingdom speakers from United States speakers with an accuracy of over 95%, and New York City speakers from other United States speakers with an accuracy of over 80%. The regression and random forest models performed similarly well. Lower performance on the New York City classification reflects the closer linguistic proximity to other American speakers compared to British speakers. Full results of all relevant models are featured in Appendix A.

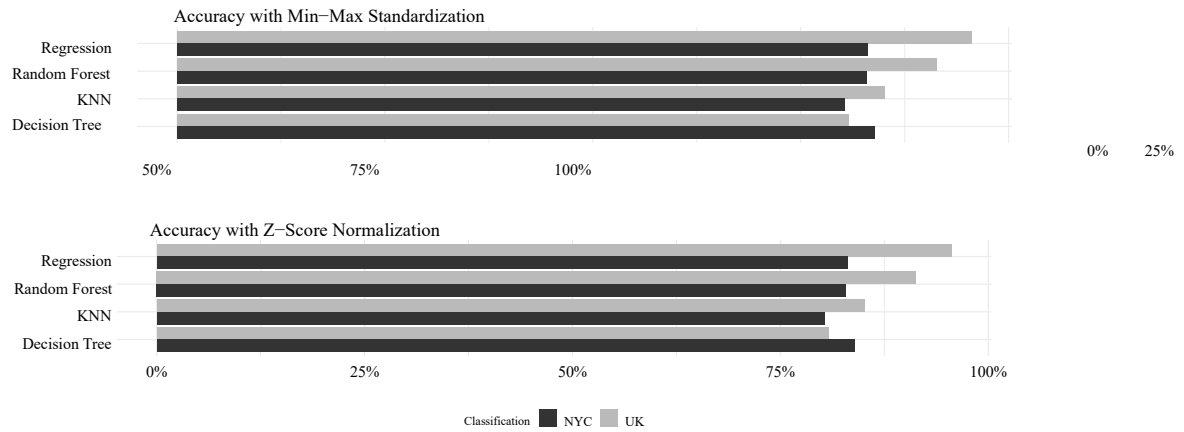


Figure 4. A comparison of various model performances. Specific values and additional comparisons can be found in Appendix A.

4.2. INTERPRETATION RESULTS. The different interpretation methods previously described each have different and unique situations that they are most applicable in. In order to account for the information encoded in each variable importance metric, an aggregation method was used to determine which phonetic features should be considered for analysis from each dataset. To be considered, a feature must have been ranked in the top 40% of features ranked by importance as quantified by regression coefficients, permutation importance, and mean decrease impurity. Features for each dataset are ranked by their mean decrease impurity scores. The top 14 features for the United Kingdom classification (Table 1) and top 10 features for the New York city classification (Table 2) as ranked and filtered by this method are focused on for analysis.

Feature	Avg USA Value	Avg UK Value	GAE	RP	Explanation
F1-also	680.8093	503.525	/ʔA:l.soU/	/ʔO:l.s@U/	cot-caught
F2-ask	1706.025	1382.939	/æsk/	/A:sk/	/æ/ raising
F1-call	686.1175	498.24	/kA:l/	/kO:l/	cot-caught
F2-bob	1214.352	1015.945	/bA:b/	/b6b/	cot-caught
F1-bob	759.4124	622.0192	/bA:b/	/b6b/	cot-caught
F2-bags	1779.109	1550.942	/bæg/	/bæg/	/æ/ raising
F2-snack	1745.38	1560.663	/snæk/	/snæk/	/æ/ raising
F1-small	707.6073	556.876	/smA:l/	/smO:l/	cot-caught
F1-frog	712.1918	601.2108	/frA:g/	/fr6g/	cot-caught
F2-slabs	1615.848	1590.662	/slæb/	/slæb/	/æ/ raising
F2-blue	1601.721	1765.391	/blu:/	/blu:/	/j/ dropping
F2-to	1637.341	1739.811	/tu:/	/tu:/	/j/ dropping
F1-bags	732.4773	767.1577	/bæg/	/bæg/	/æ/ raising
F2-spoons	1490.402	1619.809	/spu:n/	/spu:n/	/j/ dropping

Table 1. Selected UK classification features

Of the sounds identified from the classification of British and American speakers, there are three main categories that can be identified. The first is evidence of certain back vowels merging in the speech of many American speakers while the same sounds remain distinct in most British speakers. This is most prominent in words such as “call” where the vowel is pronounced like the A in “father,” with a back tongue position and less rounding (cubrovic & Bjelakovic 2020). This shift away from the shorter, rounded, and further forward /O/ sound used by most British speakers to something closer to /A/ is often referred to as the cot-caught merger (Labov et al. 2006). These features were correctly identified by the classification models due to formant values being a good acoustic representation of vowel quality. The decreased rounding and further back tongue position is acoustically represented by a substantial increase in the F1 value. The interpretation methods assigned high variable importance scores to all of the F1 values of back vowels with these properties.

The second category of sounds includes several short-a sounds such as the /æ/ in “bags.” In American speech this sound is usually pronounced with a high F2 value characteristic of most other short-a sounds. In British speech this sound is often broader, sounding much closer to the aforementioned /A/ sound of “father” (Khan & Alzobidy

Feature	Avg NYC Val	Avg Other Val	GAE Transcription	Explanation
F1-carry	657.9431	535.4434	/ʔkær.i/	short-a split
F1-all	614.0524	678.7424	/A: l/	cot-caught
F1-water	570.5861	641.6744	/ʔwA:tÄ/	cot-caught
F2-all	994.3907	1034.0973	/A:l/	cot-caught
F2-water	970.2003	1018.8752	/ʔwA:tÄ/	cot-caught
F2-had	1768.844	1839.18	/hæd /	short-a split
F1-that	736.0719	691.596	/Dæt/	short-a split
F2-carry	1764.142	1840.429	/ʔker.i/	short-a split
F1-wash	713.216	686.9142	/wA:S/	unknown
F2-dark	1304.181	1358.591	/dA:rk/	unknown

Table 2. Selected NYC classification features.

2019). The general tendency to produce these short-a sounds with a lower F2 value can be seen in Figure 5, which depicts all of the average formant measurements for each word mapped onto the vowel space. This difference can be seen in greater detail in Figure 6.

The identification of the proper name “Bob” as an important word for model classification is a somewhat special case, as the pronunciation of proper names can vary drastically from culture to culture. Nevertheless it appears that the large differences in F1 and F2 values may reflect a combination of back vowel merging and less rounding among Americans. The final category of identified features is that which is indicative of the yoddropping feature of some British dialects. In words such as “blue,” some British speakers will add an additional /j/ sound such that the pronunciation is closer to /blju/ (Khan & Alzobidy 2019). This results in an average F2 value that is higher for British speakers in general, a trait which the models use to help identify British speakers, thus making these particular words important to model performance.

These are not the only vowels sounds which differ (on average) between American speakers and British speakers. Some argue that every vowel differs in some capacity between the two regions (Khan & Alzobidy 2019), which makes the identification of the most relevant features all the more important.

Interpretation of the New York City classification identified two categories of phones as particularly important. The F1 value of “wash” and the F2 value of “dark” were both identified as weakly important, but it is not clear whether these sounds fit nearly into the known identifiers of New York City Speech. It has long been known that New York City English has a complex short-a split system (Labov et al. 2006), a characteristic not present in any other American English dialect. The system is intensely complicated, and

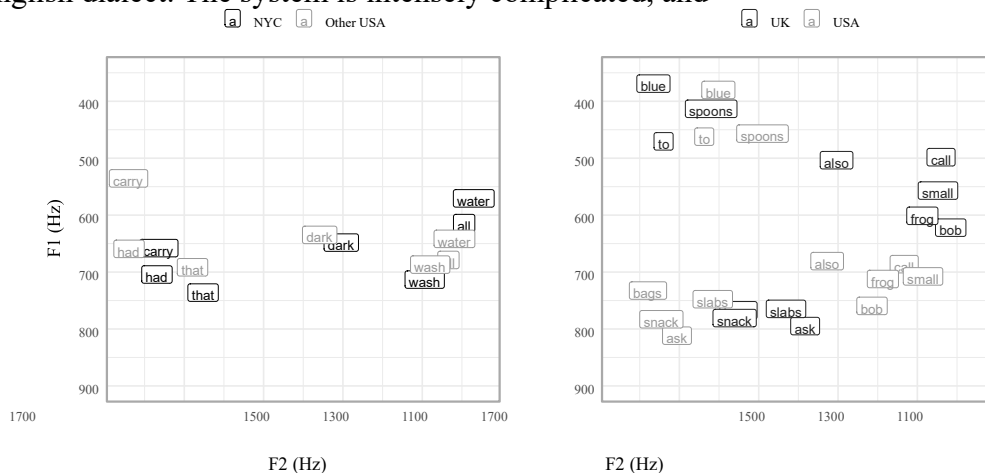


Figure 5. Words with high variable importance metrics mapped onto the vowel space. For comparison, words that were not identified as important can be seen in Appendix B.

separates particular /æ/ sounds into lax and tense varieties.

By making some of these sounds more lax, the F1 value raises, giving an acoustic quantification of this feature for the model to use when identifying New York City speakers (Newman 2014). This drift can be seen for the average speaker in Figure 5 and across all speakers in Figure 7. Additional vowel space plots demonstrating the shifts of other New York City vowels can be seen in Appendix D.

A second feature of New York City English which is much more salient to many listeners but is less unique to New York City specifically is an absence of the cot-caught merger mentioned previously. In New York City speech the /A/ sound used by many Americans in words like “call” is raised to a sound closer to /U@/, resulting in iconic New York city pronunciations such as /kU@ fi/ (Newman 2014). In model interpretation, features that have split from the /A/ sound in this way can be identified by the slight lowering of both F1 and F2 values, resulting in sounds much closer to /U/ or /O/. The vowel in “all” and the vowel in the beginning of “water” are both characteristic of this feature.

With the most important features identified by each of the pipeline demonstrations closely aligning with domain knowledge about the regional speech patterns under study, there is strong evidence that these techniques can be used to identify speech variations. With this knowledge, understudied dialects, regions, and languages can be analyzed using these techniques to better understand how the phonetics of these speakers differ from other individuals.

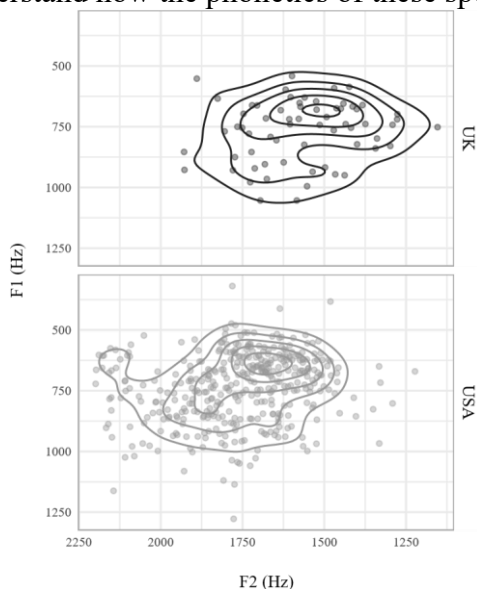


Figure 6. A comparison of UK and USA speakers’ formant measurements for the vowel in “bags.”

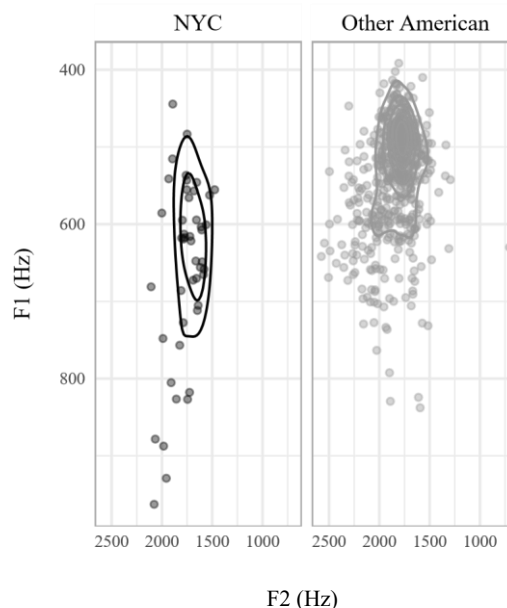


Figure 7. A comparison of NYC and other American speakers’ formant measurements for the first vowel in “carry.”

4.3. FUTURE RESEARCH AND APPLICATIONS. Further developing more advanced pipelines which better address the interactions between individual phonetic features is an obvious next step for continued research in interpretable dialect classification. Unsupervised ML algorithms which sort unlabeled utterances into categories may also benefit from interpretation methods, with the potential to expose interpretation methods to much larger datasets. In addition to these methodological advances, there are many situations and applications where these techniques could be implemented.

In traditional studies of speech variation, data is collected by establishing field sites that exist to recruit, interview, and record participants local to the area (Fridland & Kendall 2022). If this sort of system or something similar is being implemented, interpretation methods could be performed on each subregion being studied to identify which of that area’s phonetic features are

most frequently and prominently different from the remainder of the population in question. These methods could provide objective ordered lists of sounds for linguists to begin their analysis of inter-region variations. Results such as the locations of false positives predicted by particular models could provide further insights in addition to the variable importance measurements (see Appendix C).

These methods could also be used in the development of linguistic maps and atlases. Machine learning algorithms such as unsupervised cluster analyses have already been proposed for the purposes of identifying sampling sites and delineating geographic boundaries between populations of speakers (Jeszenszky et al. 2021). This supervised pipeline could supplement these tools by providing researchers with objective and thorough explanations for why speakers from different regions have been categorized into particular groups, and in what ways their phonetic inventories overlap.

5. Code availability. All code written for this paper was written in Python (processing and modeling) and R (analysis and plotting), supplemented with some Praat scripting. Code is available at <https://github.com/HenryNomeland/speech-class> along with explanations, examples, and additional plots.

References

- Baranowski, Marcin. 2022. Part of town as an independent factor: the north-force merger in manchester. *Language Variation and Change* 34(3). 239–269. <https://doi.org/10.1017/S095439452200014X>.
- Boersma, Paul & David Weenink. 2017. Praat: Doing phonetics by computer. Available online at: <http://www.praat.org/>.
- Chittaragi, N. B., A. Prakash & S. G. Koolagudi. 2017. Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering* 43(8). 4289–4302. <https://doi.org/10.1007/s13369-0172941-0>.
- Cubrovic, Biljana & Andrej Bjelakovic. 2020. Pronunciation model selection, or do you speak american? In Biljana ubrovi (ed.), *Bells90 proceedings: International conference to mark the 90th anniversary of the english department, faculty of philology, university of belgrade*, vol. 1, 139–151.
- Cychosz, Margaret. 2020. Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods* <https://doi.org/10.3758/s13428-020-01365-9>.
- D’Onofrio, Annette & Jaime Benheim. 2020. Contextualizing reversal: Local dynamics of the northern cities shift in a chicago community. *Journal of Sociolinguistics* 24(4). 469–491. <https://doi.org/10.1111/josl.12398>.
- Eide, Kristin Melum & Tor A Afarli. 2020. Dialects, registers and intraindividual variation: Outside the scope of generative frameworks? *Nordic Journal of Linguistics* 43(3). 233–248. <https://doi.org/10.1017/S0332586520000177>.
- Fridland, Valerie & Tyler Kendall. 2022. Managing sociophonetic data in a study of regional variation. *The Open Handbook of Linguistic Data Management* 237.
- Garofolo, John S. et al. 1993. *Timit acoustic-phonetic continuous speech corpus ldc93s1*. Linguistic Data Consortium Philadelphia. Web Download.
- George Mason University. 2010. Speech accent archive. <http://accent.gmu.edu/index.php>.

- Ghose, Abhishek & Balaraman Ravindran. 2020. Interpretability with accurate small models. *Frontiers in Artificial Intelligence* 3. 3. <https://doi.org/10.3389/frai.2020.00003>.
- He, Haibo & Yunqian Ma. 2013. *Imbalanced learning: Foundations, algorithms, and applications*. John Wiley & Sons. <https://doi.org/10.1002/9781118646106>.
- Hebert, Mathieu. 2008. Text-dependent speaker recognition. In *Springer handbook of speech processing*, 743–762. Springer. <https://doi.org/10.1007/978-3-540-49127-937>.
- Humayun, M. A., J. Shuja & P. E. Abas. 2024. A review of social background profiling of speakers from speech accents. *PeerJ Computer Science* 10. <https://doi.org/10.7717/peerj-cs.1984>.
- Jadoul, Yannick, Bill Thompson & Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71. 1–15. <https://doi.org/10.1016/j.wocn.2018.07.001>.
- James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani & Jonathan Taylor. 2023. *An introduction to statistical learning* Springer Texts in Statistics. Springer Cham. <https://doi.org/10.1007/978-3-031-38747-0>. <https://doi.org/10.1007/978-3-031-38747-0>.
- Jeszenszky, Pter, Carina Steiner & Adrian Leemann. 2021. Reduction of survey sites in dialectology: A new methodology based on clustering. *Frontiers in Artificial Intelligence* 4. <https://doi.org/10.3389/frai.2021.642505>. <https://www.frontiersin.org/articles/10.3389/frai.2021.642505>.
- Ke, Jiafu. 2004. *Self-organization and language evolution: System, population, and individual*: City University of Hong Kong dissertation.
- Khan, Afzal & Soleman Awad Mthkal Alzobidy. 2019. Vowel variation between american english and british english. *International Journal of English Linguistics* 9(1). 350.
- Khan, Azal Ahmad, Omkar Chaudhari & Rohitash Chandra. 2024. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications* 244. 122778. <https://doi.org/10.1016/j.eswa.2023.122778>. <https://www.sciencedirect.com/science/article/pii/S0957417423032803>.
- Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of north american english: Phonetics, phonology, and sound change: A multimedia reference tool*. Walter de Gruyter. Includes bibliographical references and index.
- Ladefoged, Peter. 1993. *A course in phonetics*. College Publishers 3rd edn.
- Lou, Zhewen & Yitian Ren. 2021. Investigating issues with machine learning for accent classification. *Journal of Physics Conference Series* <https://doi.org/10.1088/17426596/1738/1/012111>.
- McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, <https://api.semanticscholar.org/CorpusID:12418404>.

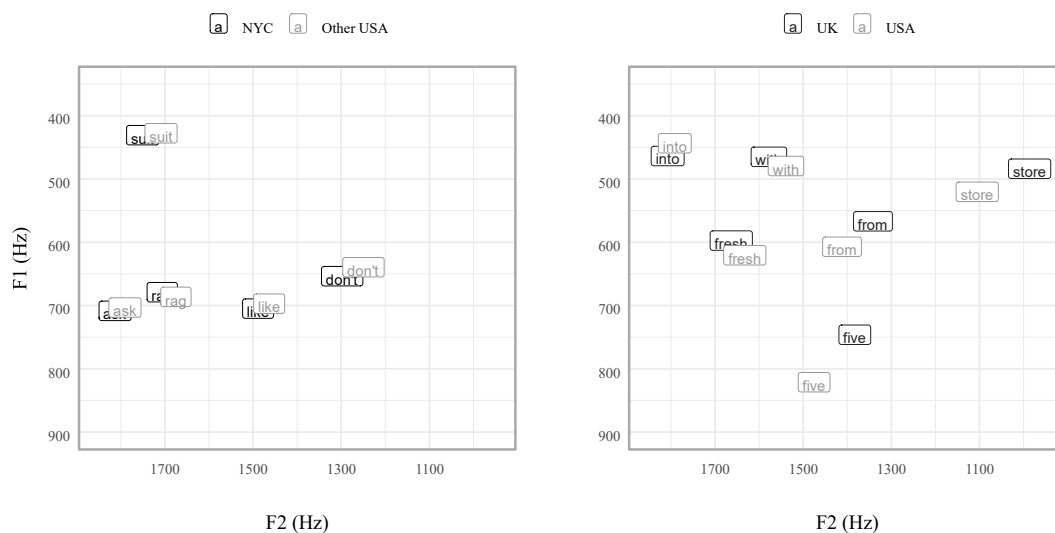
- Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl & Bin Yu. 2019. Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences* 116(44).
<https://doi.org/10.1073/pnas.1900654116>.
- Najafian, Maryam, Saeid Safavi, Peter Weber & Martin J. Russell. 2016. Identification of british english regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey*, 132–139.
- Nembrini, Stefano, Inke R Konig & MN Wright. 2018. The revival of the gini importance? *Bioinformatics* 34(21). 3711–3718.
<https://doi.org/10.1093/bioinformatics/bty373>.
- Newman, Michael. 2014. *New york city english*, vol. 10. New York City: Walter de Gruyter.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.
- Ross, Jory, Kevin D. Lilley, Cynthia G. Clopper, Jennifer S. Pardo & Susannah V. Levi. 2021. Effects of dialect-specific features and familiarity on cross-dialect phonetic convergence. *Journal of Phonetics* 86. 101041.
<https://doi.org/10.1016/j.wocn.2021.101041>. <https://doi.org/10.1016/j.wocn.2021.101041>.
- Schaeffler, Friedrich. 2005. *Phonological quantity in swedish dialects*: Ume University dissertation.
- Strycharczuk, Patrycja, Manuel Lpez-Ibez, Guy Brown & Adrian Leemann. 2020. General northern english. exploring regional variation in the north of england with machine learning. *Frontiers in Artificial Intelligence* 3. 48.
<https://doi.org/10.3389/frai.2020.00048>.
- Tanner, James, Morgan Sonderegger & Jane Stuart-Smith. 2022. Multidimensional acoustic variation in vowels across English dialects. In Garrett Nicolai & Eleanor Chodroff (eds.), *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, 72–82. Seattle, Washington: Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2022.sigmorphon-1.8>. <https://aclanthology.org/2022.sigmorphon-1.8>.
- Themistocleous, Charalambos. 2019. Dialect classification from a single sonorant sound using deep neural networks. *Frontiers in Communication* 4. 64.
<https://doi.org/10.3389/fcomm.2019.00064>. <https://doi.org/10.3389/fcomm.2019.00064>.
- Thiel, Alexander & Aaron J. Dinkin. 2020. Escaping the trap: Losing the northern cities shift in real time. *Language Variation and Change* 32(3). 373–398.
<https://doi.org/10.1017/S0954394520000137>.
- Voeten, Charlotte C. 2021. Regional variation in ongoing sound change: The case of the dutch diphthongs. *Journal of Linguistic Geography* 9(2). 162–177.
<https://doi.org/10.1017/jlg.2021.10>.

- Wei, Pengfei, Zhenzhou Lu & Jingwen Song. 2015. Variable importance analysis: A comprehensive review. *Reliability Engineering System Safety* 142. 399–432. <https://doi.org/https://doi.org/10.1016/j.ress.2015.05.018>. <https://www.sciencedirect.com/science/article/pii/S0951832015001672>.
- Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264. <https://doi.org/https://doi.org/10.1146/annurev-linguist030514-124930>.

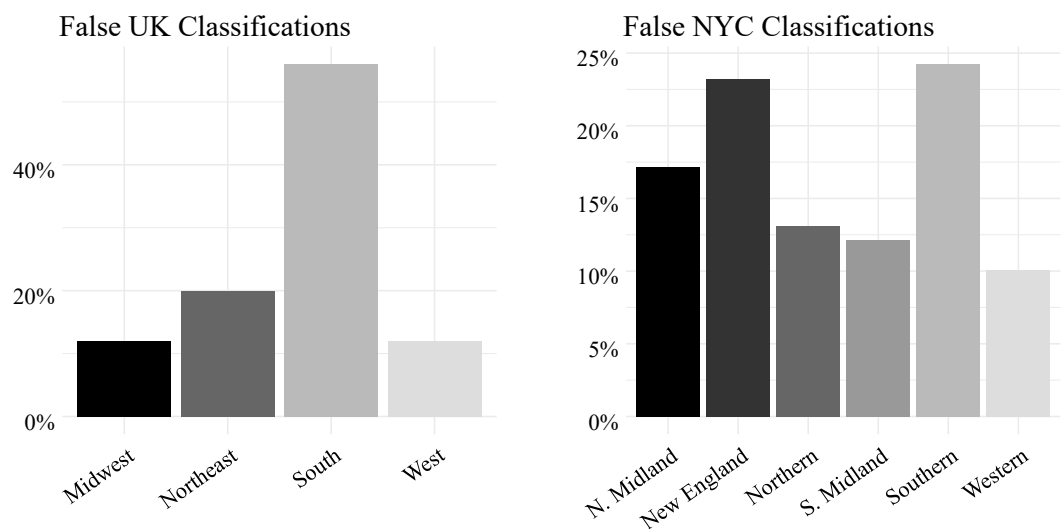
Appendix A: Model Accuracies.

Model	Standardization	Data	Accuracy
ridge regression	z	timit nyc	0.764
ridge regression	minmax	timit nyc	0.831
ridge regression	z	gmu uk	0.803
ridge regression	minmax	gmu uk	0.956
decision tree	z	timit nyc	0.732
decision tree	minmax	timit nyc	0.839
decision tree	z	gmu uk	0.77
decision tree	minmax	gmu uk	0.808
knn	z	timit nyc	0.776
knn	minmax	timit nyc	0.803
knn	z	gmu uk	0.836
knn	minmax	gmu uk	0.851
random forest	z	timit nyc	0.833
random forest	minmax	timit nyc	0.829
random forest	z	gmu uk	0.909
random forest	minmax	gmu uk	0.913
random forest	minmax	timit newengland	0.708
random forest	minmax	timit northern	0.694
random forest	minmax	timit northmidland	0.558
random forest	minmax	timit southmidland	0.687
random forest	minmax	timit southern	0.715
random forest	minmax	timit western	0.63
ridge regression	minmax	timit western	0.64

Appendix B. Low Importance Features in the Vowel Space.



Appendix C: Misclassified Sample Locations.



Appendix D: Additional Vowel Space Comparisons.

