Laboratory Phonology
Journal of the Association for Laboratory Phonology

Villarreal, D., et al. 2020 From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 6, pp. 1–31. DOI: https://doi.org/10.5334/labphon.216

JOURNAL ARTICLE

From categories to gradience: Auto-coding sociophonetic variation with random forests

Dan Villarreal[1], Lynn Clark[2,3], Jennifer Hay[2,3] and Kevin Watson[2,3]

[1] Department of Linguistics, University of Pittsburgh, Pittsburgh, PA, US

[2] New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, NZ

[3] Department of Linguistics, University of Canterbury, Christchurch, NZ

Corresponding author: Dan Villarreal (d.vill@pitt.edu)

The time-consuming nature of coding sociophonetic variables that are typically treated as categorical represents an impediment to addressing research questions around these variables that require large volumes of data. In this paper, we apply a machine learning method, random forest classification (Breiman, 2001), to automate coding (categorical prediction) of two English sociophonetic variables traditionally treated as categorical, non-prevocalic /r/ and word-medial intervocalic /t/, based on tokens' acoustic signatures. We found good performance for binary classifiers of non-prevocalic /r/ (Absent versus Present) and medial /t/ (Voiced versus Voiceless), but not for medial /t/ with a six-way coding distinction (largely due to some codes being sparsely represented in the training data). This method also yields rankings of acoustic measures in terms of importance in classification. Beyond any individual measures, this method generates probabilistic predictions of variation (classifier probabilities) that represent a composite of the acoustic cues fed into the model. In a listening experiment, we found that not only did classifier probabilities significantly capture gradience in trained listeners' perceptions of rhoticity, they better predicted listeners' perceptions than individual acoustic measures. This method thus represents a new approach to reconciling the categorical and continuous dimensions of sociophonetic variation.

Keywords: Sociophonetic variation; machine learning; rhoticity; New Zealand English

1. Introduction

Researchers in sociophonetics and variationist sociolinguistics have increasingly turned to computational methods to automate time-consuming research tasks such as data extraction (e.g., Fromont & Hay, 2012), phonetic alignment (e.g., McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017; Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), transcription (e.g., Reddy & Stanford, 2015), and measurement of vowels (e.g., Labov, Rosenfelder, & Freuhwald, 2013), consonants (e.g., Schuppler, Ernestus, Scharenborg, & Boves, 2011; Schuppler, van Dommelen, Koreman, & Ernestus, 2012; Sonderegger & Keshet, 2012), and suprasegmentals (e.g., Rosenberg, 2017). In the cases of forced phonetic alignment and automated transcription (both of which use machine learning methods), the technique rests on the assumption that there is some learnable, predictable pattern in the input that can be used to predict new cases; if an acoustic pattern surfaces in test data (data that the model was not exposed to in training) that the model has reliably seen associated with a certain label in training data, then a well-performing model will assign the same label to this pattern in the test data.

We propose that the same logic can apply to another common, time-consuming step in the process of conducting sociophonetic research: coding, or assigning categories (usually allophones), to variable data based on acoustic features. Complicating this process are two common properties of sociophonetic variables: acoustic complexity and gradient variability. When listeners (whether trained or lay) hear tokens of a sociophonetic variable, they do not only hear one or two salient acoustic measures, but rather a constellation of acoustic cues. Moreover, these cues' distributions seldom divide neatly into bins corresponding to individual allophonic variants, but are rather characterized by continuous, gradient variation; this fact belies the long-standing categorical treatment of variables like /r/, a treatment that hides from researchers' view the potential meaningfulness of tokens that exist in the gray area between cardinal variants. We thus argue that, just as human coders contend with acoustic complexity and gradient variability in making coding judgments—not always with success, as evidenced by low inter-rater reliability for variables like /r/ (see Section 3.1)—so too must an automated method for sociophonetic coding contend with acoustic complexity and gradient variability if its predictions are to have any validity.

Like the established methods of forced phonetic alignment and automated transcription, we apply a machine learning method to find learnable, predictable patterns in acoustically complex input data, in order to predict new cases—here, codes for sociophonetic variables—in test data. While several machine learning methods are likely appropriate for this application, we use the method of random forest classification (Breiman, 2001), which does not suffer from overfitting when predictors are collinear (Matsuki, Kuperman, & Van Dyke, 2016; Strobl & Zeileis, 2008); in this respect, random forests are unlike methods like generalized linear modeling (e.g., linear regression, logistic regression), for which collinearity in predictors can cause poor prediction of unseen data and misleading estimates of variable importance (Dormann et al., 2013). This property of random forests is especially important for acoustically complex variables, for which it is unclear which acoustic cues most meaningfully underlie variation; although it is not an invitation to indiscriminately include every possible acoustic parameter, this property of random forests nevertheless allows us to test large numbers of acoustic measures.

We propose that such an automated method for sociophonetic coding should differ in an important respect from the other computational methods mentioned above. In essence, those methods seek to replicate the labor of a competent human (aligning phonetic boundaries, assigning words to acoustic signals, etc.) with greater efficiency. While an automated method for coding should be able to do the same (e.g., assigning Present or Absent codes to tokens of English non-prevocalic /r/ with greater efficiency than a human could), to do so would be to ignore gradient variability. The acoustic correlates of many purportedly 'categorical' sociophonetic variables exist on a gradient, with acoustic correlates often exhibiting unimodal or overlapping distributions rather than sorting neatly into separate distributions by class (e.g., Love & Walker, 2013; see also Purse, 2019 for an articulatory case); this suggests that some variables themselves may also exist on a gradient, varying between tokens exemplary of one class to those exemplary of neither class to those exemplary of another class (Hashimoto, 2019; Hay & Clendon, 2012; Hay & Maclagan, 2010, 2012; Love & Walker, 2013). In essence, an automated method for sociophonetic coding should capture this gradient variability by learning from the acoustic complexity of training data.

Our research questions are:

1. Can we use machine learning methods to effectively code sociophonetic variables (in particular, New Zealand English non-prevocalic /r/ and word-medial intervocalic /t/)?
2. If machine learning methods can effectively code these sociophonetic variables, how much data is necessary to make it work?

3. How well do classifiers' gradient predictions predict trained listeners' coding judgments?

In the following sections, we present an automated method for sociophonetic coding that successfully meets the desiderata outlined above. We first provide a background on random forests and classification (Sections 2 and 3). Section 4 is devoted to describing the classifiers, including the results (Section 4.3) and simulations to explore the relationship between the size of the training data and classifier performance (Section 4.5). In Section 5, we describe a listening experiment that assessed the accuracy of the /r/ classifier's predictions and the degree to which the /r/ classifier's gradient predictions matched trained listeners' coding judgments. We conclude with a general discussion in Section 6. Readers can find additional information in the online supplementary materials, which detail the hyperparameter settings used in the classifiers and give additional information about simulations described in Sections 4.3.2 and 4.5. Readers can also find a step-by-step guide to training random-forest classifiers (including all the data for both variables in this paper, and all the code needed to recreate the /r/ classifier) at https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html.

2. Random forests and classification

Random forests are an extension of the method of classification and regression trees, which recursively partition data into successively smaller subsets at each tree node by finding the independent variable that minimizes variation in the subbranches under the node (Breiman, 2001). While individual trees are useful for describing variation in the contexts of classification (categorical dependent variable) or regression (continuous dependent variable), they suffer from overfitting to training data and so perform poorly at predicting unseen data. Random forests are an ensemble method that grow a multitude of systematically 'weaker' trees—which are only permitted to select from among random subsets of independent variables at each split—that together form a consensus on the classes of previously unseen observations and which predictors are most important (Tagliamonte & Baayen, 2012). This method reduces variance in estimates without the overfitting that individual stronger trees suffer from (Breiman, 2001). Importantly, unlike modeling techniques like generalized linear modeling, random forests do not suffer from overfitting when predictors are collinear (Dormann et al., 2013; Matsuki et al., 2016; Strobl & Zeileis, 2008); as a result, collinearity does not hinder random forests' ability to predict unseen data and to determine the relative importance of independent variables.[1] This property is particularly important for the present study, as the acoustic complexity of /r/ and /t/ led us to include many acoustic measures, including those that are naturally correlated with one another (e.g., formant values at multiple timepoints), in our random forests; indeed, we demonstrate in Section 4.3.2 that our data is characterized by a considerable degree of collinearity. We should note here that other classification methods

[1] In generalized linear modeling, all predictors are evaluated simultaneously; thus, if two predictors are collinear, they share information about the dependent variable, so it is difficult to separate their relative contributions to overall variance. Furthermore, if the collinearity structure of unseen data is different from that on which a generalized linear model is trained, prediction of unseen data fails spectacularly with even mild collinearity (Dormann et al., 2013). Random forests, by contrast, do not suffer the effects of collinearity on either prediction or variable importance: "Due to the randomly restricted variable selection scheme employed in random forests even variables that would be considered redundant by most regression approaches can receive an equally high importance score, because the different randomly restricted tree models in the ensemble can reflect the effects of a variable in different contexts with different covariates" (Strobl & Zeileis, 2008, p. 2). Indeed, random forests performed toward the top among the 23 methods evaluated by Dormann et al.'s (2013) collinearity simulations, especially with respect to the severe collinearity that likely characterizes acoustic datasets.

have this property, and we do not claim that random forests are the only approach to automatically coding sociophonetic variation; as the goal of the present research is to present one method for automated sociophonetic coding and not to compare methods, we welcome future research that looks more closely into the benefits and drawbacks of different machine-learning tools for this application.

In this paper, we use random forests for three specific functions: classification (predicting the response values of observations in a test set based on the patterning of predictors and responses in a training set), feature selection (determining which predictors are most influential in classification), and gradience (mapping categorical classes to continuous predictions). (Again, these functions are not exclusive to random forests.) The classification function has been used in fields like photogrammetry (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012) and sleep medicine (Fraiwan, Lweesy, Khasawneh, Wenz, & Dickhaus, 2012) but not, to our knowledge, in phonetics or sociolinguistics. In a related approach, German, Carlson, and Pierrehumbert (2013) used optimal discriminant analysis to calculate speaker-specific F3 thresholds for distinguishing retroflex versus flapped realizations of English /r/.

The feature selection function of random forests has been used in several studies of cue weighting (often with acoustic cues). For example, Brown, Winter, Idemaru, and Grawunder (2014) explored which of nine phonetic cues were most predictive of Korean honorific judgments for both Korean and English listeners; Al-Tamimi (2017) contrasted 13 acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic; and Baumann and Winter (2018) explored 17 prosodic and non-prosodic cues affecting perceptions of prominence. In a slightly different function, Tagliamonte and Baayen (2012) used random forests to compare the importance of linguistic and extralinguistic factors on *was/were* variation in York English, finding that individual variation outranked all other predictors in importance.

Finally, in the gradience function of machine-learning models like random forests, a model trained on categorical data generates continuous, probabilistic predictions, which can potentially be used to predict listener responses on new data. This function is related to our earlier observation that the acoustic correlates of many purportedly 'categorical' sociophonetic variables exist on a gradient (see Section 1); if the continuous nature of these correlates indeed accrues to the gradience of the variables themselves—which is by no means a given—then these models can predict this gradience. For example, van Alphen and Smits (2004) use individual regression trees to predict the likelihood that listeners hear tokens of Dutch labial and alveolar plosives as voiced (although they do not apply these probabilistic predictions to new tokens).

Within sociolinguistics, we are aware of only one study that attempts to use the gradience function of machine-learning models; McLarty, Jones, and Hall (2019) train a support vector machine (SVM) classifier of African American Language non-prevocalic /r/s. Using 36 mel-frequency cepstral coefficients as acoustic measures, the authors obtained binary classifications and continuous measures of how far tokens were from the decision boundary for non-prevocalic /r/s (defined as the vowel + /r/ sequence, as in the present research). However, their classifier was trained on oral vowels (representing the Absent class) and onset /r/s (representing the Present class). We note that this approach rests on the premise that the acoustic continuum of variation in non-prevocalic /r/s is identical to the continuum of variation between onset /r/s and vowels; there are several reasons to doubt this premise. First, even a fully constricted non-prevocalic /r/ will take up less of the time-course of its token than an onset /r/; this dissimilarity between onset and Present coda /r/ is problematic given the coarse temporal resolution—three time points—of the acoustic measures fed into this classifier. Second, there is no oral vowel analogue for the

Absent versions of contexts like SQUARE; it is not clear how well a classifier whose training set of Absent tokens does not include [ɛə] could classify such tokens as Absent. We are skeptical that a classifier whose training and test sets differed in such a fundamental way could succeed, regardless of the specific machine-learning implementation used, the size of the training set, or the number of acoustic measures. Because the authors neither report cross-validation results on non-prevocalic /r/ data with known codes nor compare the classifier's predictions against human coders, it is not possible to validate the predictions of the classifier that they report.

3. Background: /r/ and /t/

Our classifier efforts focused on two sociophonetic variables: non-prevocalic /r/ and medial /t/. Both of these variables are prime candidates for this method: Variants of these variables are sociolinguistically meaningful in New Zealand English (NZE) (and numerous Englishes worldwide), making the automation of coding advantageous for research; and they are acoustically complex, meaning for neither variable is there a single acoustic measure that can stand in as a proxy for its variability. In this section, we discuss each variable's acoustic features and sociolinguistic context in NZE, as well as our corpora.

3.1. Non-prevocalic /r/

The typical approach to non-prevocalic /r/ (aka postvocalic /r/, rhotic /r/) in sociolinguistics is to treat /r/ as binary, varying between Present and Absent (or rhotic and non-rhotic) (Bartlett, 2002; Becker, 2009; Gordon et al., 2004; Labov, Ash, & Boberg, 2006; Nagy & Irwin, 2010). This distinction conflates several phonetic processes, depending on lexical set: coda [ɹ] versus compensatory lengthening (e.g., START [stɑɹt~stɑːt]), coda [ɹ] versus a centering offglide (e.g., NEAR [nɪɹ~nɪː]), and r-coloring versus no r-coloring (e.g., NURSE [nɜ˞s~nɜːs]).[2] As we discuss below, inter-rater reliability is notoriously low for /r/, even among trained phoneticians (e.g., Yaeger-Dror et al., 2009).

A strand of sociophonetic research treats /r/ as continuous by measuring the F3 minimum of the /r/ token, with lower F3 minimum corresponding to a greater constriction and thus a 'stronger' /r/. Common to this research is evidence for treating /r/ as meaningfully gradient. Working with a very small sample from the Mobile Unit of early NZE, Hay and Clendon (2012) find that speakers with higher rhoticity levels also had lower F3 values in the tokens that were analyzed as having /r/ Present. In other words, speakers with more /r/ tokens, also had more Present-like tokens (see also Hay & Maclagan, 2010 for intrusive /r/). While the above analyses focused on speakers, Hay and Maclagan (2012) also considered variation in the F3 of words produced with word-final linking /r/ (where the /r/ surfaces across a word boundary before a vowel). Some word tokens were associated with a high probability of occurring prevocalically and thus a high probability of attracting an /r/. The authors show a correlation between the likelihood that a word occurs prevocalically and the F3 of /r/s produced in that word. In other words, words that were more likely to occur with an /r/ were also more likely to contain a more Present-like /r/. Hashimoto (2019) finds analogous results for Māori loanwords borrowed into NZE (e.g., *kia ora* 'hello'). Some proportion of the time, these loanwords are produced with a tap sound, imported from Māori; on other occasions they are produced with an adapted approximant /r/. Hashimoto shows that the more often a loanword is produced with an approximant /r/, the lower the F3 of the approximant is likely to be. Finally, Love and Walker (2013) found that British English speakers' /r/s had lower F3s (i.e., more Present-like) when discussing American football than when discussing soccer, but even these more

[2] Here, we use Wells' (1982) lexical set notation.

Present-like /r/s had F3s well above that of American English speakers' /r/s. In summary, the axis of variation in /r/ is not limited to the difference between Present and Absent tokens but phonetic differences within and between these classes.

While F3 minimum is clearly relevant to /r/ variation, several acoustic studies have found a complex array of cues to rhoticity: a greater lag between F3 minimum and the end of the token (Lawson, Stuart-Smith, & Scobbie, 2018), longer duration and lower F2 (Stuart-Smith, 2007), spectral information below F3 (Heselwood, 2009), and a lack of frication noise immediately after the token (Stuart-Smith, Lawson, & Scobbie, 2014). Complicating matters further is that rhoticity subsumes distinct articulatory strategies (Lawson, Scobbie, & Stuart-Smith, 2014), with acoustic ramifications for retroflex versus bunched /r/ (Zhou et al., 2008). As a result, it is not clear that F3 on its own is sufficient for modeling continuous variation in /r/.

Rhoticity existed at low levels in the early history of NZE, especially among speakers born before 1875 and in majority-Scottish settlements (Gordon et al., 2004), but has gradually declined and disappeared in General NZE (Hay & Sudbury, 2005). Some literature suggests the recent re-emergence of a rhotic /r/, particularly in Auckland and Northland and as a marker of Māori/Pasifika identity (Gibson, 2005; Kennedy, 2006). Southland New Zealand English is widely identified by New Zealanders as the country's only regional dialect (e.g., Nielson & Hay, 2005). The variety is spoken in the country's southernmost region, which has a heavily Scottish settlement history. The primary linguistic feature that sets it apart from General NZE is that it has remained at least partially rhotic throughout its history, especially following the NURSE vowel (Bartlett, 2002). As described in 4.1 below, the data for building our /r/ random forest classifier came from a corpus of Southland English.

3.2. Medial /t/

Word-medial intervocalic /t/ is also subject to complex variation in NZE and other varieties, as it can undergo a multitude of phonological processes that include frication, voicing, flapping, and glottalization; Hay and Foulkes (2016) report at least ten variants. Unsurprisingly, the acoustic cues associated with variation in medial /t/ (and English /t/ in other contexts; Temple, 2014) are likewise complex; frication is associated with a greater center of gravity, spectral dispersion, and kurtosis (Jones & Llamas, 2008; Jones & McDougall, 2009), and shorter duration (Buizza & Plug, 2012); voicing is associated with the presence of periodicity (Riehl, 2003); flapping with weakened F2 and F3 (Warner & Tucker, 2011); and glottalization with decreased formant transitions (Docherty & Foulkes, 1999) and increased spectral tilt (Seyfarth & Garellek, 2015). Within sociolinguistics, to simplify this complex variation and facilitate statistical modeling, medial /t/ variants are sometimes collapsed into binary categories. For example within NZE, several studies have modeled a Voiced versus Voiceless distinction (Clark, 2018; Hay & Foulkes, 2016) to capture one major axis of ongoing change. In NZE, the conservative variant [t] has moved toward two gender-marked variants: The move toward [ɾ] is led by men (Hay & Foulkes, 2016; Holmes, 1994), and a contemporaneous shift toward [s] (particularly in formal contexts) is led by women (Fiasson, 2015).

3.3. Summary

Both non-prevocalic /r/ and word-medial intervocalic /t/ are variables which are often subject to manual coding for sociolinguistic variants. This is time consuming, and also potentially obscures important dimensions of acoustic variability. At least for /r/, the coding may itself be influenced by the coder's own expectations and previous experience of rhoticity (Hay, Drager, & Gibson, 2018; Yaeger-Dror et al., 2009), making the categorical coding of /r/ particularly problematic. Indeed, /r/ is notoriously difficult to code reliably, even among trained phoneticians; Lawson et al. (2014) coded Glaswegian English /r/ into seven categories, and allowing for one category leeway (e.g., counting 'no /r/' and 'derhotic' as agreement), the three authors achieved 84–86% agreement. Comparable findings for other variables have been reported by Irwin (1970) (86–87% test-retest consistency for labeling consonant misarticulations); Pitt, Johnson, Hume, Kiesling, and Raymond (2005) (92.9% inter-coder agreement for stops, 86.5% for liquids); Fosler-Lussier, Dilley, Tyson, and Pitt (2007) (79.4–80.9% inter-coder agreement for stops, 74.7–79.0% for liquids/glides); and Hall-Lew and Fix (2012) (by-token SDs around 0.8–1.0 for coding /l/ vocalization on a 4-point scale). Hall-Lew and Fix (2012) also note that the 'intermediate' tokens (those with mean ratings in the middle of the scale) experienced the least reliable ratings.

Because these variables are common targets of time-consuming manual sociolinguistic coding, and because they vary in interesting ways both across dialects and across time within dialects, we chose them as useful targets to explore the value of automated coding. We hoped to be able to train automatic classifiers to create large data-sets of automatically coded sociolinguistic variants. Because these variables are also characterized by acoustic complexity, we hoped to use these classifiers to investigate the acoustic measures that correspond to categorical variation in these variables. We further hoped to investigate classifier probability as a useful proxy for phonetic gradience—one that simultaneously takes into account multiple phonetic cues.

4. Building classifiers

We trained random forest classifiers in an attempt to automate coding for our two sociophonetic variables: non-prevocalic /r/ and medial /t/. Using training sets of several thousand hand-coded tokens (/r/: 4,689; /t/: 4,218), we trained classifiers on large numbers of acoustic measures (180 measures of /r/, 113 measures of /t/). Our question is whether this method is viable for automatically coding these variables. In this section, we first outline the data used for this automatic classification, before outlining the acoustic measures that were extracted from each token and the performance metric that we used to tune the classifiers.

4.1. Training and test data

Data for the /r/ classifier came from our corpus of Southland English, which consists of over 83 hours of mostly spontaneous speech recordings of 113 Southland English speakers born between 1868–1998. This corpus is hosted in an instance of LaBB-CAT (Fromont & Hay, 2012), a browser-based linguistics research tool that stores audio and/or video recordings, and text transcripts which have already been time-aligned at the utterance level (e.g., using software such as ELAN, Sloetjes & Wittenburg, 2008). LaBB-CAT then allows these files to be automatically force aligned to the word and segment level with the Hidden Markov Model Toolkit (Young et al., 2006); these alignments can then form the basis of batch phonetic measurements using Praat (Boersma & Weenink, 2015). The New Zealand Institute of Language, Brain and Behaviour curates several different spoken corpora using LaBB-CAT and some of these already contain speakers from Southland. Our Southland corpus was partly compiled by migrating those existing recordings into a new instance of LaBB-CAT. This was then supplemented with new data given to us for research purposes by the Southland Oral History Project and Invercargill City Libraries.

The /r/ data was partitioned into a *training set* (tokens with known codes on which the classifier was trained and tuned) and a *test set* (uncoded tokens on which the final classifier was applied) based on which tokens had already been hand-coded. Of the 5,901 total hand-coded tokens that we had access to, the final training set comprised 4,689 tokens (see Section 4.2.1); the test set comprised 27,516 additional tokens, and stimuli in the experimental study in Section 5 were drawn from this test set.[3] The /r/ training data came from 28 sociolinguistic interviews conducted for Chris Bartlett's (2002) doctoral thesis corpus, recorded in Southland in 1992 with speakers born between 1904–1977. These interviews were recorded in mono at a sampling frequency of 22,050 Hz and saved as wav files before being uploaded to our Southland LaBB-CAT corpus. All hand-coding was performed by Chris Bartlett, who coded tokens into binary Present versus Absent; tokens coded as Absent (72.2%) well outnumbered Present (27.8%). Test data was drawn from the rest of the Southland corpus.

The /t/ data for training and testing came from QuakeBox, a corpus of 512 speakers' narratives, recorded in 2012, about their experiences of the 2010–11 Canterbury earthquakes (Clark, MacGougan, Hay, & Walsh, 2016). In particular, we drew 4,218 hand-coded tokens from 168 QuakeBox speakers, representing over 29 hours of talk. These narratives were recorded in stereo at a sampling frequency of 48,000 Hz and saved as wav files. A research assistant hand-coded medial /t/ into six classes, [ɾ s d t ʔ ∅].[4] In this training set, two majority classes ([ɾ s]) were considerably more prevalent than the others: [ɾ] 39.2%; [s] 39.0%; [d] 9.5%; [t] 7.2%; [ʔ] 3.3%; [∅] 1.8%. We trained a classifier on the medial /t/ data with all six classes. When this six-class classifier failed to perform up to standards (Section 4.3), we collapsed these classes to a binary Voiced versus Voiceless distinction (following Clark, 2018; Hay & Foulkes, 2016): Voiced 50.5%; Voiceless 49.5%. This binary classifier was then used to predict 4,888 additional /t/ tokens from 207 QuakeBox speakers, which had not been previously hand-coded.

4.2. Methods

4.2.1. Acoustic measures

For both variables, we extracted a large number of acoustic measures to serve as training data for the classifiers. Our choice of measures was guided by two aims: producing a well-performing classifier, and resolving persistent uncertainty about the acoustic features that best characterize variation in /r/ and /t/. The latter aim meant that we did not directly introduce any social factors or linguistic factors above the level of phonetics (e.g., speaker gender, preceding vowel, stress) into the classifiers; we also wanted to avoid the classifiers over-learning extra-phonetic features, as these features may have applied in different degrees to the training versus test sets.[5] This second aim is also why we did not pursue features like mel-frequency cepstral coefficients that are popular in the fields of signal processing and speech recognition (e.g., Wei, Cheong-Fat, Chiu-Sing, & Kong-Pang,

[3] Machine learning literature sometimes uses the term *test set* to refer to data that is held out during training for the purposes of assessing the classifier's performance, for example in *k*-fold cross-validation; if this held-out set is used for model selection, it is often referred to as a *validation set* (Hastie, Tibshirani, & Friedman, 2009, p. 222). In this case, since we apply the classifier to data for which the 'ground truth' is unknown, we reserve the term 'test set' for this previously uncoded data. When discussing partitioning of the training set in the online classifier-training tutorial (https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html), we refer to training and test *subsets*.

[4] This data was previously analyzed in Clark (2018).

[5] As an example of overlearning, consider a training set that comes from speakers among whom the strongest predictor of rhoticity is preceding vowel, with far greater rhoticity after NURSE than in other vowel contexts. Now suppose that the test set comes from speakers for whom preceding vowel is less relevant as a constraint on rhoticity (i.e., who exhibit equivalent levels of rhoticity after NURSE as they do elsewhere). A classifier trained on this training set's acoustic measures *and* phonological properties (namely, preceding vowel) will erroneously predict greater rhoticity after NURSE in the test set, even in the absence of evidence from the acoustics of the test set of greater rhoticity after NURSE. In other words, the introduction of extra-phonetic features raises the risk of the classifier basing its predictions not on acoustics, but on features that may be in flux under conditions of intra-community change (e.g., Dodsworth & Kohn, 2012) or cross-community diffusion (e.g., Nagy & Irwin, 2010). In fact, this flux in extra-phonetic features characterized the speech community from which we drew /r/ data; across the 20th century, Southland English went from being variably rhotic in all preceding vowel contexts to rhoticity being limited to after NURSE.

2006), opting instead for measures that have more currency in acoustic phonetics (cf. McLarty et al., 2019). Acoustic measure extraction was performed in LaBB-CAT (Fromont & Hay, 2012) via batch-application of measurements in Praat (Boersma & Weenink, 2015).

For /r/, we extracted 180 measures corresponding to potential acoustic cues of rhoticity (see Section 3.1):

- F1, F2, F3, and F4 (speaker-normalized) at 13 timepoints
- Formant maxima and minima (speaker-normalized) and the normalized times at which maxima and minima were found
- Formant ranges (raw maximum minus raw minimum), and slopes (range divided by normalized time)
- Differences between raw formant values at 13 timepoints (e.g., differences between F2 and F1, F3 and F2)
- Formant bandwidths at 13 timepoints
- Pitch maxima and minima, normalized timepoints of maxima and minima, pitch range, and pitch slope
- Amplitudes at F3 maxima and minima
- Token duration, *z*-scored by speaker

Due to some measurement error in automatically extracting formant and pitch measurements, and to account for the degree to which formant frequencies are impacted by vocal tract length, we subjected /r/ measurements to pre-processing before entering them into the classifier. These pre-processing steps for the /r/ measurements are explained in greater detail in our online classifier-training tutorial (https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html). In brief, we:

1. excluded 131 tokens (2.2%) that had bad duration measurements or preceded another /r/
2. normalized formant frequency measurements based on speakers' initial /r/s (see e.g., Hay & Clendon, 2012)
3. centered and scaled token duration by speaker and vowel
4. imputed missing F0 and F4 measurements (494 tokens imputed, 8.7%)
5. excluded 28 tokens (0.5%) with missing measurements other than F0 or F4, and
6. excluded 931 tokens (16.6%) for which there were outliers in selected measures, as determined by preliminary testing: F3 at the 80% timepoint, F2 at the 35% timepoint, F1 at 50% and 70%, F3 minimum, and intensity at the F3 minimum timepoint.

For /t/, we extracted 113 measures (see Section 3.2):

- Amplitudes at 13 timepoints
- Change in amplitude between token onset and minimum amplitude
- Center of gravity at 13 timepoints
- Dispersion at 13 timepoints
- Kurtosis at 13 timepoints
- Voicing at 13 timepoints (a binary TRUE/FALSE based on whether Praat's autocorrelated pitch reading was defined or undefined)
- Spectral tilt of the 25 ms immediately preceding the token[6]

[6] The Praat code to extract spectral tilt was based on a script by Christian DiCanio: http://www.acsu.buffalo.edu/~cdicanio/scripts/Get_Spectral_Tilt_2.praat.

- F1, F2, and F3 bandwidths at 13 timepoints
- F1, F2, and F3 formant transitions (the difference between formant measurements at the token boundary and 25 ms away from the token boundary, for both left and right boundaries)
- Token duration, *z*-scored by speaker

Unlike the /r/ data, the distributions of /t/ measures featured far less measurement error or missing measurements, so we did not impute any missing measurements in the /t/ data or exclude any tokens for having outliers. As in the /r/ data, we standardized token duration by speaker. We excluded 208 tokens (4.7%) that had missing measurements.

4.2.2. Performance metric: Overall accuracy × AUC

We subjected classifiers to a process of hyperparameter tuning to find the classifier that maximized our chosen performance metric: area under the ROC curve (AUC) × overall accuracy. Overall accuracy is simply the percentage of correct predictions; in automated coding of sociophonetic variation, as in hand-coding, we want to maximize the percentage of tokens that are coded accurately. However, overall accuracy is not sufficient to capture the quality of a classifier by itself. This principle can be illustrated by a hypothetical /r/ data set in which both training and test sets are 93% Absent, 7% Present. Suppose we train a classifier on the training set and it predicts that the test set is 100% Absent, 0% Present. This classifier thus achieves an impressive overall accuracy of 93%, but with an alarming disparity between the Absent class accuracy (93%) and the Present class accuracy (0%); we would gladly sacrifice some of the Absent class accuracy if we knew our classifier didn't wildly under-predict Present, and our performance metric should reflect that tradeoff.[7] In short, overall accuracy is insufficient to capture classifier performance, especially when classes are imbalanced in size and thus class accuracies are likely to differ (as in the /r/ data).

AUC, on the other hand, quantifies the ability of a classifier to balance true positives and true negatives at the set of all possible classification thresholds (Hastie et al., 2009). For example, the /r/ classifier estimated each test set token's probability of being Present, which we label *classifier probability*, and converted classifier probabilities to binary codes by specifying a threshold above which tokens would be coded Present. This threshold defaults to 0.5, but it is possible to measure the Absent and Present class accuracies at different threshold values and plot these rates as a curve (the receiver operating characteristic curve, ROC) in [0,1] × [0,1] space; the area under the ROC curve (AUC) ranges from [0,1] and represents how well the classifier can balance class accuracies with one another.[8] Within this set of possible classifiers (differing only by thresholds), we choose the threshold that maximizes overall accuracy. Our custom performance metric, overall accuracy × AUC, thus rewards classifiers for high overall accuracy while penalizing classifiers that fail to adequately balance true positives and true negatives; the naive classifier described in the previous paragraph, which failed to balance Present and Absent predictions, would be discarded on the basis of having poor overall accuracy × AUC.[9]

[7] Here, we distinguish 'class accuracy,' the percentage of (e.g.) actual Absent tokens correctly classified as Absent, from 'overall accuracy,' the percentage of all tokens that were correctly classified (i.e., the average of all class accuracies, weighted by the prevalence of each class in the data). In different methodological traditions, class accuracy is also known as 'sensitivity/specificity,' 'true positive/negative rate,' or 'recall,' depending on which class is identified as 'positive.'

[8] The abbreviation AUC is also used for the area under the precision–recall curve, but here we use it strictly for the area under the ROC curve.

[9] An anonymous reviewer asks why we do not use a more widespread performance metric like F1 for optimizing classifiers. F1 is problematic for sociophonetic coding because F1 considers only true positives, false positives, and false negatives, and not true negatives; it is thus sensitive to the choice of one class as the

We ran classifiers in R (version 3.5.2) using the packages ranger and caret (Kuhn, 2018; R Core Team, 2018; Wright & Ziegler, 2017), on a computing cluster running Ubuntu 16.04 with 32 virtual cores and 64 GB of memory. The online supplementary materials include details about the hyperparameter values that emerged from the tuning process (number of trees, number of variables tested at each node, splitting rule, minimum node size, training/test subset generation method, subset generation parameters, and additional resampling). All of the /r/ and /t/ data, and all of the code necessary for running the /r/ classifier—as well as detailed explanations of these hyperparameters—can be found at https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html.

4.3. Results

In this section, we discuss the final classifiers' performance, variable importance, and auto-coding.

4.3.1. Classifier performance

Table 1 displays performance measures for the classifiers. The overall accuracy for the /r/ classifier, 84.5%, compares favorably with human coders' inter-rater reliability for coding rhoticity, as noted in Section 3.3. It is also worth noting that the class accuracy for Present lagged Absent, as the classifier correctly coded 62.2% of actual Present tokens as Present, compared to correctly coding 93.1% of actual Absent tokens as Absent. This result is not surprising given that Absent tokens outweighed Present by roughly 2.6 to 1, as the classifier had fewer Present tokens in its training set.

The six-class medial /t/ classifier achieved high class accuracies for the majority classes [ɾ s] but dismal accuracies for the minority classes [d t ʔ ∅]; in fact, none of actual [∅] tokens were correctly coded as [∅]. This classifier's confusion matrix (**Table 2**) shows that the minority classes were consistently miscoded as the majority classes (e.g., most [d] were miscoded as [ɾ], most [t] as [s]), indicating that the sizes of these classes contributed to the classifier's performance in identifying them. Clearly, this six-class classifier would be inadequate for deploying across a corpus, especially if it is important to capture rare but sociolinguistically salient variants. For example, [t] is a conservative variant in

Table 1: Classifier results. Baseline performance metrics reflecting a naive classifier (one always choosing the majority class) are given in parentheses; the .5000 baseline for AUC assumes performance equal to a coin flip.

Performance measure	/r/ classifier	Six-class /t/ classifier	Two-class /t/ classifier
Overall accuracy × AUC	.7423 (.3608)	.7107 (.1962)	.8918 (.2527)
Overall accuracy	.8447 (.7217)	.7847 (.3924)	.9182 (.5055)
AUC	.8786 (.5000)	.9056 (.5000)	.9711 (.5000)
Class accuracies	Absent 93.1%; Present 62.2%	[ɾ] 91.9%; [s] 92.7%; [d] 20.1%; [t] 37.3%; [ʔ] 47.5%; [∅] 0.0%	Voiced 92.2%; Voiceless 91.5%

'positive' class. In the naive classifier example, assuming a sample size of 100, F1 calculated from the perspective of the majority Absent class as the 'positive' class is .9637, but from the perspective of the minority Present class F1 is 0. While F1 may be useful for classification applications such as spam detection where there is a clear 'positive' class, in sociophonetic coding no class is more important for purposes of detection; a metric like overall accuracy × AUC, which is symmetric with respect to the choice of 'positive' class, is thus more appropriate. While there do exist other symmetric performance metrics such as the Matthews correlation coefficient (see e.g., Chicco, 2017), we find overall accuracy × AUC to have a straightforward, intuitive interpretation.

Table 2: Confusion matrix for six-class medial /t/ classifier (percentages of overall data). Actual classes in columns, predicted classes in rows.

	ɾ	**s**	**d**	**t**	**ʔ**	**∅**
ɾ	36.2	2.1	6.4	0.3	1.0	1.4
s	1.9	36.2	0.4	3.7	0.2	0.2
d	0.7	0.2	1.9	0.5	0.3	0.1
t	0.2	0.5	0.5	2.7	0.2	0.0
ʔ	0.4	0.0	0.3	0.0	1.5	0.0
∅	0.0	0.0	0.0	0.0	0.0	0.0

NZE, so it is useful to differentiate [t] from other variants to measure the time-course of sociolinguistic change in a corpus, but less than 40% of hand-coded [t]s were successfully detected by the classifier.[10] The poor performance of this six-class classifier motivated the attempt to treat medial /t/ as a two-class variable (following Clark, 2018; Hay & Foulkes, 2016).

By contrast, the two-class /t/ classifier performed very well, with nearly identical class accuracies for Voiced versus Voiceless. Again, to the extent that classifier performance on a given class is related to the size of that class, this result is not surprising, as the Voiced and Voiceless classes were nearly equal in size. However, this increased performance comes at the cost of failing to capture rare but sociolinguistically salient variants like [t], as with the six-class /t/ classifier.

4.3.2. Acoustic measures

Figure 1 displays the 20 most important acoustic measures for the /r/ and two-class /t/ classifiers, in order of decreasing importance, as measured by the Gini index (see the online classifier-training tutorial for more details).[11] For /r/, nearly all of these measures involved the difference between (raw) F3 and F1 (11 measures) or F3 and F2 (6 measures). **Figure 2** demonstrates that F3 – F1 was most important just after the midpoint, whereas F3 – F2 was most important toward the end of the token. Also represented in the most important measures for /r/ were token duration, intensities at F3 minimum and F3 maximum, and speaker-normalized F3 minimum.

For the two-class /t/ classifier, the most important acoustic measures likewise cluster toward the end of the token, including spectral moments (center of gravity: 8 measures; kurtosis: 4) and voicing (7); speaker-normalized token duration is also among these top 20 measures. These measures indicate that for this variable, the phonological categorization of these classes as 'voiced' and 'voiceless' (Clark, 2018; Hay & Foulkes, 2016) is only partially rooted in phonetic voicing (periodicity); instead, Voiced and Voiceless variants are distinguished by spectral characteristics that characterize fricatives.

As mentioned in Section 2, random forests do not suffer from overfitting when predictors are collinear (Matsuki et al., 2016; Strobl & Zeileis, 2008). As a check on collinearity, we calculated pairwise correlations for the acoustic measures in **Figure 1**. Out of the 190 pairwise correlations per variable, 64 correlations for /r/ and 49 for /t/ exceeded the

[10] In light of the fact that, for reasons of sociolinguistic salience, the most interesting contrasts were among [ɾ s t], we also tuned a four-class classifier that collapsed [d ʔ ∅] into an 'other' category. This classifier actually performed worse than the six-class version, with an overall accuracy × AUC of .6654, so we did not pursue this approach further.

[11] Note that while it is valid to compare variable importance scores' ranking between classifiers, variable importance scores' magnitudes should not be compared between classifiers, as scores' magnitudes are partially a function of tuning parameter settings (Strobl, Malley, & Tutz, 2009).

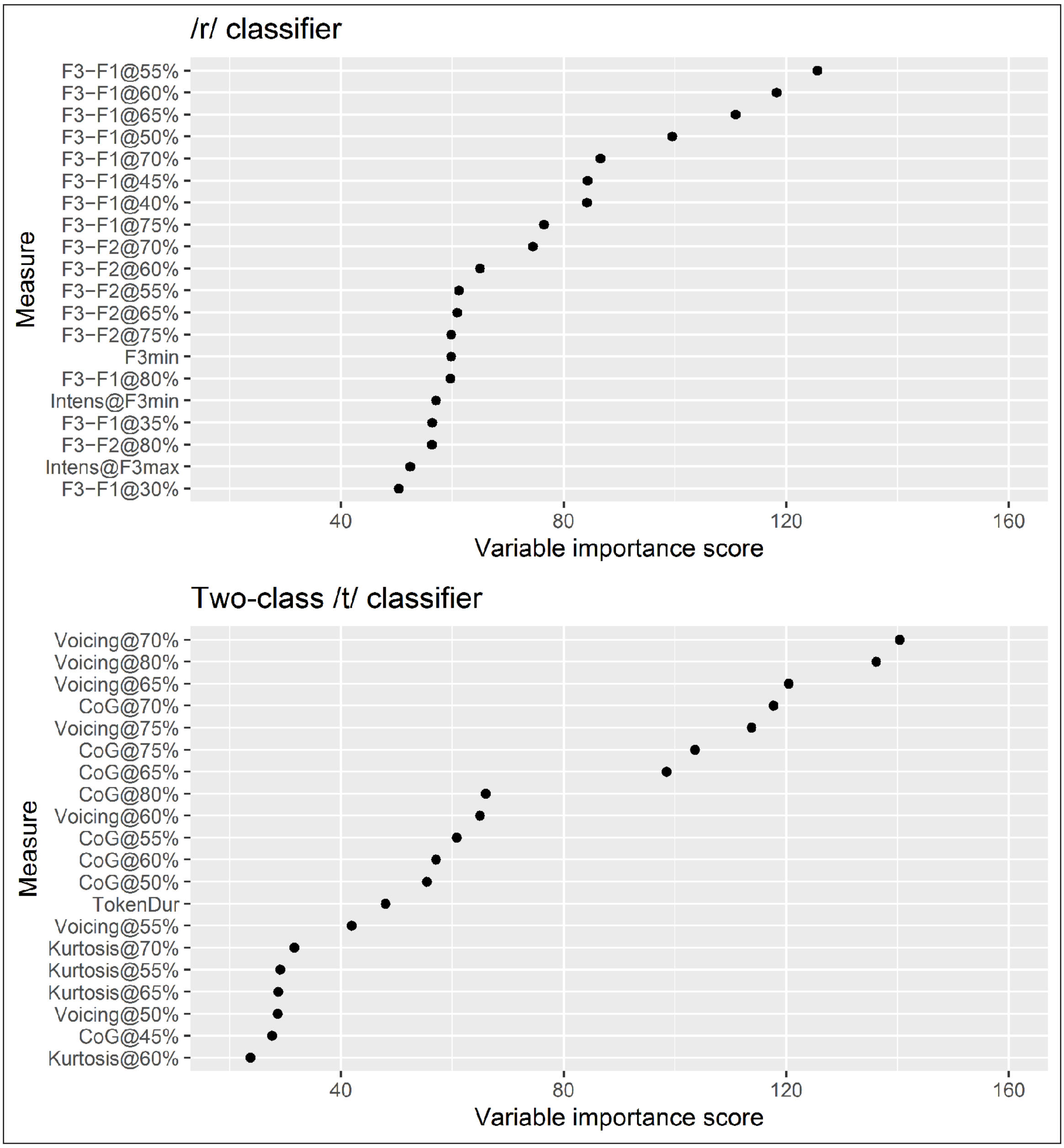

Figure 1: Top 20 most important acoustic measures (Gini index) for the non-prevocalic /r/ and two-class medial /t/ classifiers.

0.7 rule-of-thumb threshold for 'severe' collinearity (Dormann et al., 2013). A technique like generalized linear modeling that is sensitive to collinearity would clearly not be appropriate for this data.

Moreover, there is some evidence that this technique produces reliable measures of variable importance. In simulations with subsets of the /r/ and two-class /t/ data, we found high correlations between the variable importance scores (*r*s > .7728), indicating stability between estimates of variable importance (even with much smaller datasets). (Details of these simulations can be found in the online supplementary materials.[12]) This result validates our use of random forests for sociophonetic applications of feature selection, such as determining which acoustic features are most influential in classifying variants of sociophonetic variables like /r/ and /t/.

[12] Thanks to an anonymous reviewer for suggesting this analysis.

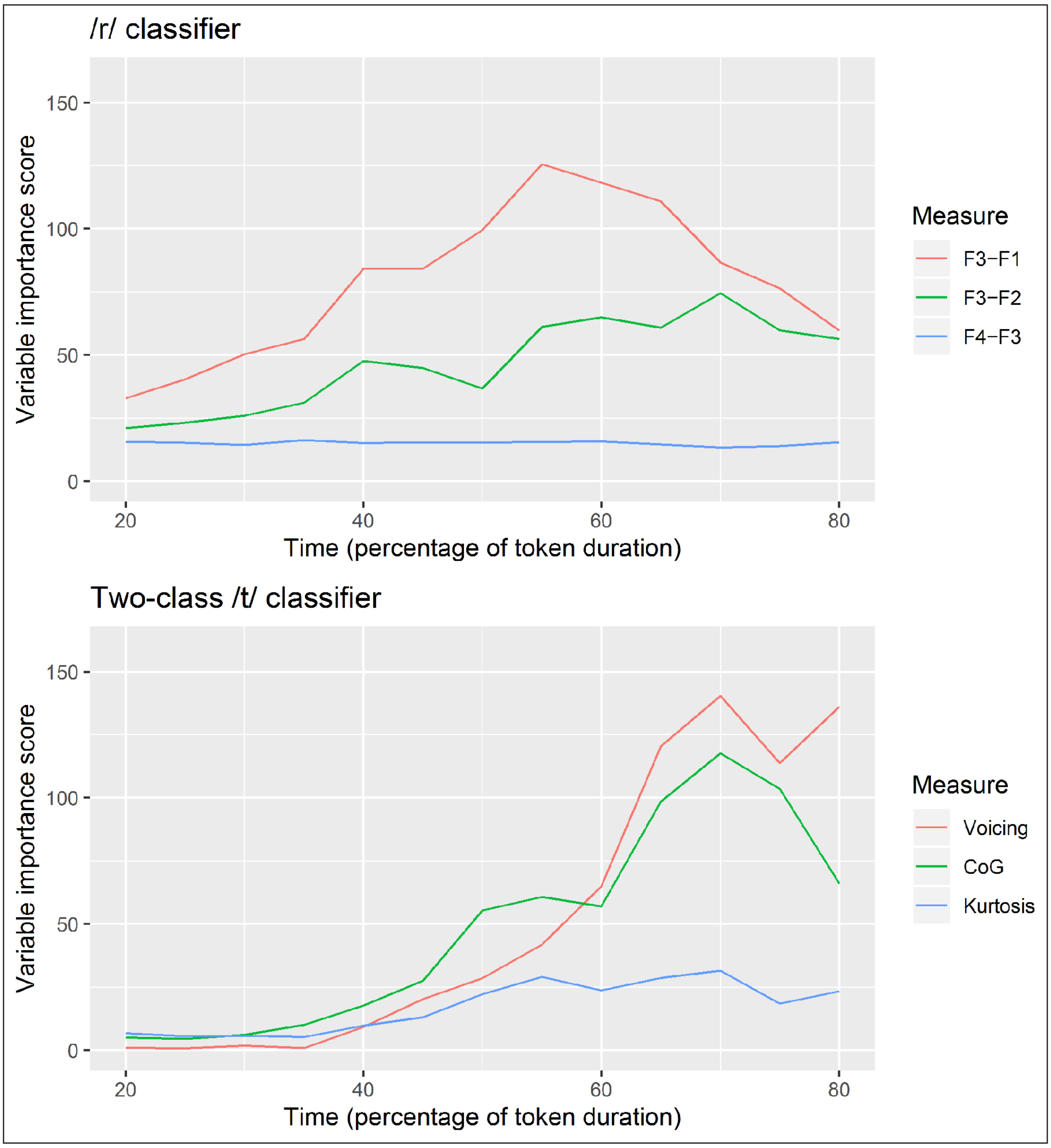

Figure 2: Variable importance for selected groups of time-varying acoustic measures for the non-prevocalic /r/ and two-class medial /t/ classifiers.

4.3.3. Auto-coding data

After training the classifiers, we used them to predict binary classes and classifier probabilities for 27,516 /r/ tokens and 4,888 /t/ tokens that had not been previously hand-coded. In terms of binary predictions, the predicted distribution of /t/ in the auto-coded test set (Voiced 52.6%; Voiceless 47.4%) closely matched the observed distribution in the /t/ training set (Voiced 50.5%; Voiceless 49.5%), whereas the predicted distribution of /r/ in the auto-coded test set (Absent 79.5%; Present 20.5%) was somewhat more skewed toward Absent than the observed distribution in the training set (Absent 72.2%; Present 27.8%). **Figure 3** displays the distributions of classifier probabilities underlying these predictions. For both variables, the distributions are similar for training and test sets, with a lower (more Absent-like) mode in the /r/ training set than the /r/ test set that may indicate greater differentiation between Absent and Present classes in the training set than the test set. However, there is a noticeable difference between the unimodal distributions for /r/ and the sharply bimodal distributions for /t/.

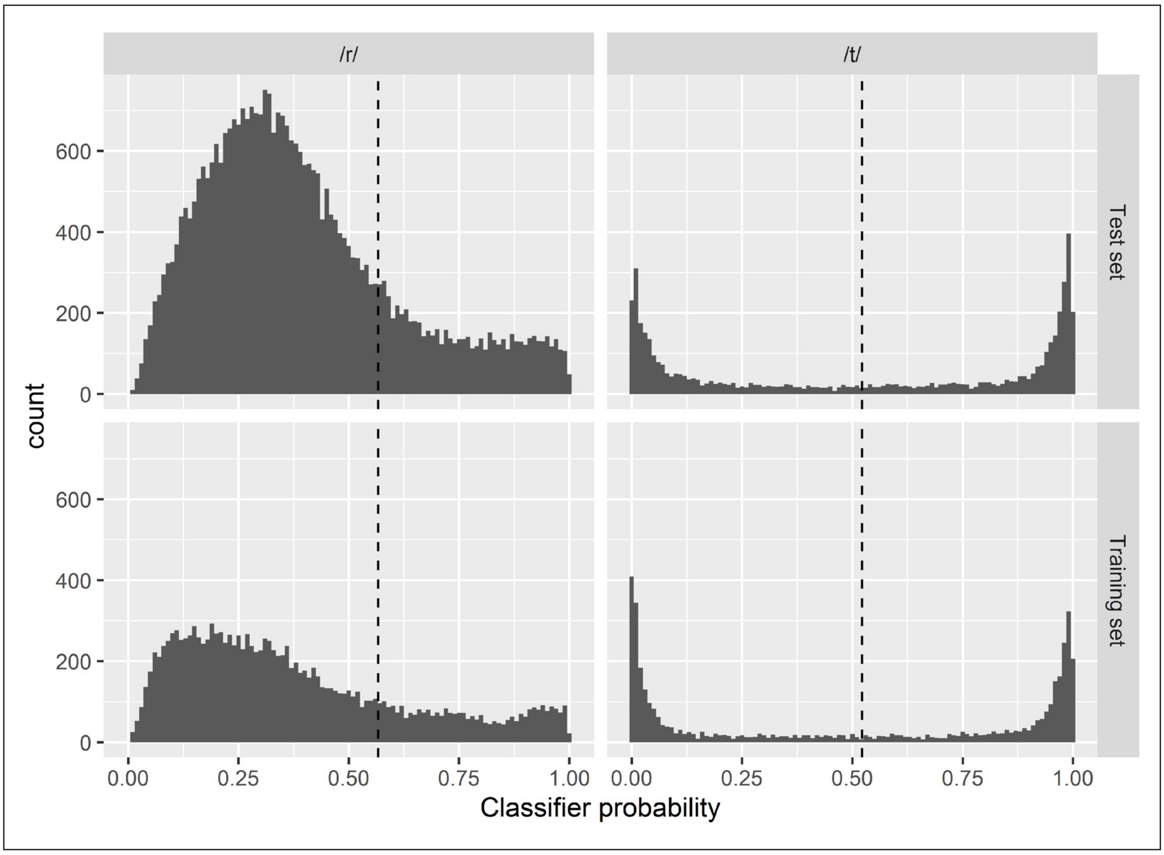

Figure 3: Distribution of classifier probabilities in training set and test set, for /r/ and two-class /t/ classifiers (greater classifier probability = more Present and more Voiced, respectively). Dotted vertical lines denote binary classification cutoffs for each classifier.

We present these comparisons between the classifier probability distributions of training sets and test sets with a caveat: Classifier probabilities are generated via slightly different processes for training versus test sets. For example, the /r/ classifier used repeated *k*-fold cross validation, meaning the test set is exposed to the entire classifier whereas the training set is exposed to just a subset. As a result, while there is no reason not to combine hand-coded and auto-coded binary codes, we caution against commingling training set and test set classifier probabilities in the same data.

4.4. *Discussion*

The performance results for the binary classifiers (for non-prevocalic /r/ and medial /t/) stand as a proof of concept for using a random forest classifier to automatically code unseen data. (We explore a different means of validating classifier performance, comparing predictions of *unseen* data to human listeners' judgments, in Section 5.) Both classifiers achieved overall accuracy rates that rival inter-rater reliability for human listeners' coding of acoustically complex variables such as these. The high AUC for both classifiers indicated a good ability to balance class accuracies; this is especially the case for the two-class /t/ classifier. In addition, this method represents a useful approach to better understanding acoustically complex variables like /r/ and /t/; by ranking acoustic measures in terms of importance for differentiating variants, this approach reveals acoustic structure that hasn't previously been accounted for. For example, whereas previous research has often treated F3 minimum as a continuous index of rhoticity (e.g., Hay & Maclagan, 2012; Love & Walker, 2013), our classifier reveals F3 minimum to be overshadowed in importance by numerous other measures (all of which relate to F3) in coding /r/ as Present or Absent.[13] Moreover, the random forest approach is particularly useful in this context; whereas collinearity hinders prediction and feature selection with modeling techniques like generalized linear modeling, random forests perform well on prediction and feature selection even under conditions of collinearity (Dormann et al., 2013; Matsuki et al., 2016; Strobl et al., 2008). This property of random forests (and other machine-learning methods) allows us to be inclusive with the acoustic measures that we feed into classifiers, even if those measures are naturally correlated with one another (although of course this property is not license to indiscriminately include every possible acoustic parameter in the model).

Whereas the binary /r/ and /t/ classifiers performed well enough to function as autocoders, the six-class medial /t/ classifier performed poorly—its overall accuracy was under 80%, and for four out of six classes, the model correctly predicted the variant in less than half of tokens—such that it would be inadvisable to deploy on a set of unseen data. There are several possible explanations for this performance deficit. One option is that the number of classes was too high to effectively cover all of them in predictions; this explanation is unlikely, however, given the existence in other domains of random forest classifiers with more classes with very good performance, such as classifiers of handwritten numerals (e.g., Zamani, Souri, Rashidi, & Kasaei, 2015). A second possibility is that the minority classes were simply too small, meaning the classifier was unable to construct an acoustic template for finding new examples of these classes in test data; under this explanation, a minority class would need to have well more than 400 tokens (the number of [d] tokens) to have a chance at being classified accurately. A third possibility is that there is a tradeoff between minority class size and consistency; in other words, a minority class that is homogeneous can achieve a greater class accuracy with a smaller sample size than a minority class that is heterogeneous. Supporting this third explanation is a result from a four-class version of the /t/ classifier that we trained in which we collapsed [d ʔ ∅] into an ‘other’ category. Although this ‘other’ category had 2,193 tokens, it had a poor class accuracy (46.0%); it is possible that since this category combined variants with rather different acoustic properties, 2193 tokens were insufficient to train the classifier to recognize the ‘other’ class. We explore these possibilities in the following section.

4.5. How much data is necessary?

We conducted an analysis of sample sizes to explore how the size and consistency of the training set affects the performance of the classifier. Practically speaking, if a researcher wants to use a random forest classifier on sociophonetic variation in their own data, they will need to know how much data to hand-code in order to train a classifier that will yield reliable predictions on non-hand-coded data. We carried out this analysis via simulations in which we re-ran the final /r/ and binary /t/ classifiers at sample sizes between 100 and 1,200 (with 100 random subsamples per sample size). Details of the implementation of these simulations can be found in the online supplementary materials. While it is impossible to come up with a ‘magic number’ minimum sample size that will work for all variables and any research question, we present these results as a general guide to future users of the classifier method.

As **Figure 4** indicates, some classifiers trained on as few as 100 tokens matched or even eclipsed the performance of the classifiers trained on the full data sets of over 4,200 tokens. As sample size increases, not only does performance improve on average, performance becomes more reliable. Improvement is more pronounced for the /t/ classifiers than the

[13] The fact that many of the most important measures are differences between F3 and F1 or F2 may also reflect that formant measures were incompletely normalized.

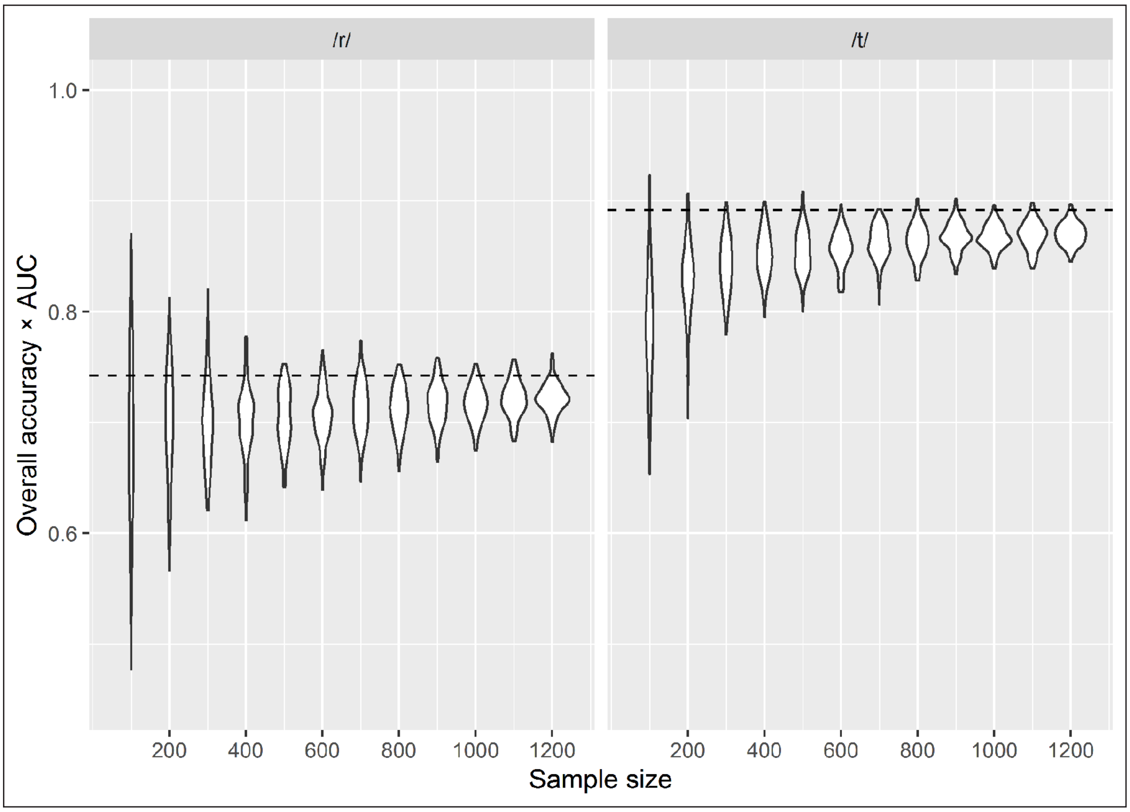

Figure 4: Overall accuracy × AUC for /r/ and /t/ simulations, 100 classifiers at each sample size. Dotted line marks performance of best classifier for each variable.

/r/ classifiers. However, for both /r/ and /t/ the average performance of classifiers run on 1,200 tokens sits well below the performance of the classifiers run on over 4,200 tokens, suggesting that more data is a sure way to improve classifier performance (albeit with diminishing returns).

Additional simulations of /r/ controlling for whether the Present and Absent classes are internally homogeneous or heterogeneous furthermore suggest that the homogeneity of the training set is associated with classifier success (plots in the online supplementary materials). When the minority Present class is homogeneous (i.e., characterized by less acoustic variability), greater classifier performance and greater class accuracies can be achieved with smaller samples. This finding suggests that the classifier has an easier time discriminating between classes when there is less variation in one class.

5. Experimental assessment of classifier predictions

In order to assess the accuracy of the /r/ classifier's predictions, and also the degree to which continuous variability in classifier probability might be sociolinguistically meaningful, we conducted a small-scale perception task. We chose to investigate the /r/ classifier and not the /t/ classifier because of the apparent gradience of /r/ and categoricity of /t/; moreover, the scarcity of /t/ tokens with intermediate classifier probabilities would make stimulus selection difficult.

Stimuli were selected from among classifier-coded tokens to represent a range of classifier probabilities and to control for additional independent factors that affect /r/ in this community. Preliminary mixed-effects modeling of production in Southland /r/ (including both hand-coded and auto-coded tokens) uncovered several significant linguistic and social factors; this analysis indicated that experimental stimuli needed to be controlled for gender, stress, morphological status, preceding vowel, and following segment. Rather than generate a complex model with interactions of all possible control factors, we chose to set each factor constant. We thus restricted our stimuli to tokens uttered by men in stressed syllables in content words, with preceding NURSE and a following sonorant. To control for any effects of metrical structure, we selected only monosyllables. Out of 27,516 classifier-coded tokens, this gave us a stimulus population of 330. Sixty tokens were then selected to span the range of classifier probabilities, with 31 tokens coded Absent and 29 coded Present by the classifier. Stimuli were created by extracting the relevant word from the audio file, resampling the word to 22,050 Hz, and scaling the word's intensity to 70 Pa.

Eleven phonetically trained listeners were asked to judge each stimulus as Present or Absent and to rate their confidence on a scale from 1 to 5. Listeners heard each stimulus word twice, with a 750 ms buffer between repetitions. Listeners performed the task with headphones on individual computers. All listeners were proficient English users who lived in a predominantly non-rhotic country at the time of the experiment. Five listeners self-reported speaking English with rhotic accents, six with non-rhotic accents; seven listeners self-reported as native speakers, four as non-native speakers.

5.1. Analysis

Prior to modeling, we calculated each listener's proportion of Present judgments ($M = 50.4\%$, $SD = 19.0\%$). We planned to discard the data of any listeners whose proportion of Present fell outside the range $M \pm 2\ SD$; despite a considerable amount of variation in proportions of Present (ranging from 13.3%–86.7%) all 11 listeners fell within this range, so no responses were discarded. (This wide range in individual responses was also observed by Yaeger-Dror et al., 2009.) The analysis sample included 659 observations (with one stimulus accidentally skipped by one listener).

Mixed-effects modeling was performed to examine to what extent classifier probability predicted listener judgments and confidence. Modeling was performed in R (R Core Team, 2018); judgments data (binary) were modeled via logistic regression in the lme4 package (Bates, Mächler, Bolker, & Walker, 2015), and confidence data (continuous) via linear regression with Satterthwaite approximations for degrees of freedom in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016; Satterthwaite, 1946). For both judgments and confidence, the baseline model included classifier probability and following segment (/l, m, n/) as fixed effects, with random effects for listener and stimulus and random slopes of classifier probability by listener. For judgments, these slopes yielded singular fits but we retained them to prevent anti-conservative *p*-values (Barr, Levy, Scheepers, & Tily, 2013). As a result, in all judgments models we removed correlations between listener intercepts and by-listener classifier probability slopes, which did not significantly affect model fit (see e.g., Sonderegger, Wagner, & Torreira, 2018); the fixed-effects model coefficients in this model were virtually identical to that without the random slope. For confidence (but not judgments, as these led to singular fits), the stimulus random effect was nested within word.

We suspected that the relationship of classifier probability to judgments might exhibit nonlinearities, so we also modeled classifier probability as a restricted cubic spline with three knots in the R package rms (Harrell, 2018). We further tested models adding listener rhoticity or native speakerhood as a main effect and as an interaction with classifier probability. These models were compared to the baseline via likelihood ratio tests. Finally, to assess the utility of classifier probability as a composite index of rhoticity above and beyond the individual measures, we ran models of judgments with individual acoustic measures as predictors and compared these to models with classifier probability via likelihood ratio tests.

5.2. Results

5.2.1. Judgments

The baseline model of judgments revealed a significant positive effect of classifier probability on trained listeners' judgments ($\beta = 3.64$, $z = 5.64$, $p < .0001$), indicating that stimuli with greater classifier probabilities were more likely to be judged Present; this relationship is evident in **Figure 5**. Following segment was not significant (*ps* $> .56$). This baseline model proved to be the best model of human judgments. The model with a restricted cubic spline for classifier probability failed to significantly improve the model fit ($\chi^2[1] = 0.10$, $p = .76$). Models that added listener rhoticity ($\chi^2[2] = 1.45$, $p = .48$) or listener native speakerhood ($\chi^2[2] = 2.80$, $p = .25$) also failed to significantly improve fit. While these nonsignificant results nominally fail to indicate either nonlinearities in the relationship between classifier probability and human judgments, or an effect of these particular dimensions of listener experience on judgments, they must be interpreted with caution; the relatively small sample size and small number of listeners means that the models have low power, especially when testing for effects of properties of speakers (Brysbaert & Stevens, 2018; Judd, Westfall, & Kenny, 2017; Kirby & Sonderegger, 2018; Westfall, Kenny, & Judd, 2014).[14]

5.2.2. Confidence

The model of confidence ratings revealed no significant main effect of classifier probability on confidence ratings, with the intercept term corresponding to the Absent judgment ($\beta = -0.16$, $t[33.43] = -0.44$, $p = .66$); however, there was a significant interaction

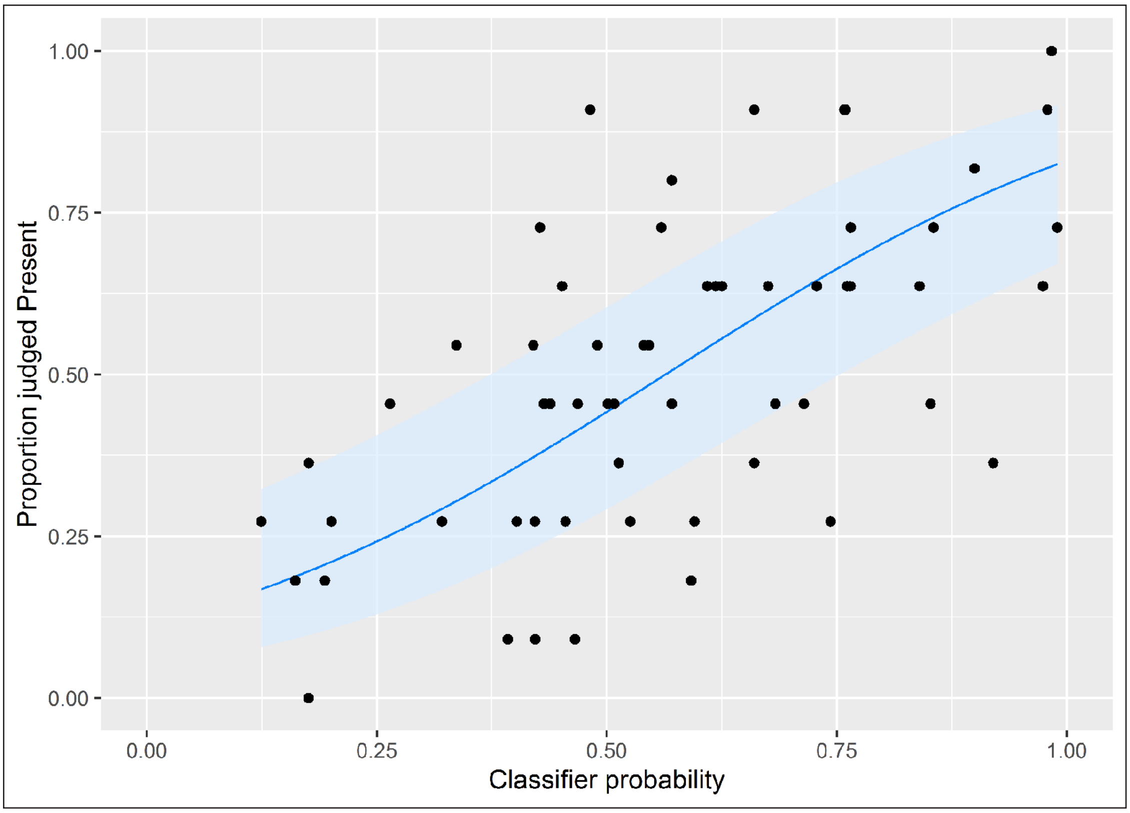

Figure 5: Classifier probability versus trained listeners' coding judgments for each stimulus (dots); fitted effects line and 95% confidence-interval bands calculated by R effects package (Fox & Weisberg, 2019) from best model of judgments.

[14] Thanks to an anonymous reviewer for making this point.

effect of classifier probability with the Present judgment (β = 1.78, $t[462.58] = 4.09$, $p < .0001$). In other words, listeners were more confident in judging stimuli with greater classifier probabilities as Present, but there is no evidence that classifier probability had an effect on listeners' confidence in Absent judgments. The model additionally revealed a significant main effect of Present judgments (β = −0.88, $t[476.83] = -3.26$, $p < .005$), indicating that listeners were less confident in judging tokens with low classifier probabilities Present than Absent. These relationships are evident in **Figure 6**.

As with judgment, the baseline models of confidence revealed no significant effect of following segment on confidence (*ps* > .24). The baseline models were also the best models of confidence, as adding a restricted cubic spline for classifier probability, listener rhoticity, and listener native speakerhood failed to significantly improve the baseline confidence model (*ps* > .69). As with the judgment models, these results nominally fail to indicate nonlinearities in the relationship between classifier probability and confidence, or that these particular dimensions of listener experience affected confidence ratings—but we again urge caution in not over-interpreting these null results given the likelihood of low power in this model (Brysbaert & Stevens, 2018; Judd et al., 2017; Kirby & Sonderegger, 2018; Westfall et al., 2014).[15]

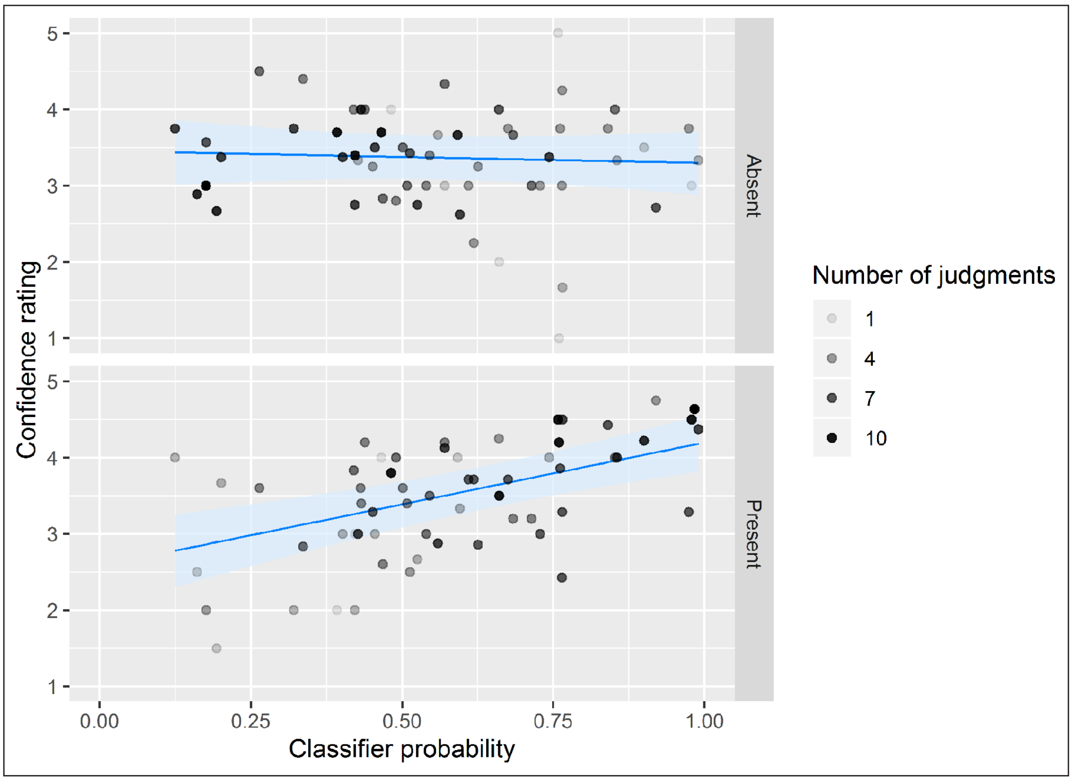

Figure 6: Classifier probability versus mean confidence rating for stimuli judged Absent versus Present (darker dots received more judgments); fitted effects lines and 95% confidence-interval bands calculated by R effects package (Fox & Weisberg, 2019) from best model of confidence.

[15] We fit the same models of confidence as ordinal regression models (in particular, cumulative link mixed models) using the ordinal package for R (Christensen, 2019). The results were qualitatively similar. The baseline model had no significant main effect of classifier probability, a significant main effect of the Present judgment, a significant interaction effect between classifier probability and Present judgments, and no main effect of following segment; as with the linear models, in the ordinal models, adding a restricted cubic spline for classifier probability, listener rhoticity, and listener native speakerhood failed to significantly improve the baseline confidence model.

5.2.3. Classifier probability versus individual measures

As mentioned in Section 3.1, sociophonetic modeling of /r/ as a continuous phenomenon (i.e., beyond the Present/Absent binary) typically treats F3 minimum as an index of rhoticity (Hay & Maclagan, 2012; Love & Walker, 2013). It is an empirical question, then, whether F3 minimum significantly predicts human judgments in the same way that classifier probability does—and if so, which measure predicts human judgments better. To test this question, we ran six additional models of judgments: three models with an individual acoustic measure (centered and scaled) instead of classifier probability, and three with both classifier probability and an individual measure. Two measures were F3 minimum: the raw F3 minimum, and the speaker-normalized F3 minimum that was entered into the classifier. Since the latter was ranked only 14th out of 180 variables in importance by the classifier, we also assessed the top-ranked variable in importance (see Section 4.3.2): the difference between F3 and F1 at 55% of the token's duration. To parallel the models reported in Section 5.2.1, these models included random slopes of the acoustic measure by listener that were uncorrelated with the listener random intercept, even when the correlated slopes did not result in a singular fit; in no case did removing this correlation significantly hurt the model fit.

In all of these models, the individual measure significantly or near-significantly affected trained listeners' coding judgments in the expected (negative) direction (**Table 3**); for example, tokens with greater raw F3 minimum were significantly less likely to be judged Present. However, likelihood ratio tests revealed that adding classifier probability to these models resulted in significantly better fits; for example, the model with raw F3 minimum plus classifier probability was significantly better than the model with just raw F3 minimum ($\chi^2[2] = 9.27, p < .01$).[16] In other words, classifier probability captured something about human judgments above and beyond what these individual measures can capture.

We should note that the stimulus sample was chosen to systematically represent a range of classifier probabilities, while we did not specifically sample for F3 minimum and other individual measures. To control for this, we resampled the stimuli such that the shape of the classifier probability distribution more closely matched the shape of the raw F3 minimum distribution, and we repeated the above comparison with models of this resampled data set. The models with classifier probability and individual measures were significantly better than the models with just individual measures for normalized F3 minimum ($\chi^2[2] = 9.22, p < .01$) and F3 − F1 at 55% duration ($\chi^2[2] = 24.16$, $p < .0001$), but not for raw F3 minimum ($\chi^2[2] = 4.79, p = .09$).

Table 3: Significance of main predictors (centered and scaled) in individual-measure judgment models. "Log-lik 1/2" refers to log-likelihoods of models without/with classifier probability, respectively.

Main predictor	**Estimate**	**Log-lik 1**	**Log-lik 2**	**Likelihood ratio test**
Raw F3 minimum	$\beta = -0.80, z = -4.39, p < .0001$	−386.87	−382.24	$\chi^2(2) = 9.27, p < .01$
Normalized F3 minimum	$\beta = -0.67, z = -3.73, p < .001$	−390.68	−382.52	$\chi^2(2) = 16.33, p < .001$
F3–F1 at 55% duration	$\beta = -0.45, z = -2.68, p < .01$	−395.35	−383.04	$\chi^2(2) = 24.62, p < .0001$

[16] Collinearity is a natural concern for these models given that these individual measures are part of the composite that makes up classifier probability. We found that the variance inflation factors (VIFs) of these models, as calculated by the R package car (Fox & Weisberg, 2019), were well below 4, the most conservative of typical VIF thresholds (O'Brien, 2007).

5.3. Discussion

Predictions from a random forest classifier trained on binary coded data can accurately predict gradient responses of phonetically trained listeners in a binary-coding task. This gradience is seen in two ways. First, while listeners varied substantially in the likelihood of hearing an /r/, the classifier was able to predict their behavior as a group (**Figure 5**). And second, when individual listeners did hear an /r/ as Present, the classifier was able to predict how confident they were in that judgement. The latter was not true when they coded the /r/ as Absent. That is, listeners tend to hear different degrees of /r/ presence more clearly than they hear different degrees of /r/ absence.

These results suggest several important implications for sociolinguistic and phonetic studies of /r/. First, they shed some light on the acoustic complexity of /r/ by indicating that classifier probability outstrips individual cues in predicting listeners' judgments. The difference between F3 and F1 shortly after the midpoint emerged as the cue that contributed to classifier performance most. While judgments were significantly related to this cue as well as raw and normalized F3 minimum, these models of individual measures were all significantly improved-upon with the addition of classifier probability. The percept of an /r/ is almost certainly influenced by a conglomerate of acoustic properties—and no individual property may be reliably present across all tokens—though it is not a given that classifier probability would successfully represent such an acoustic conglomerate. While our trained listeners vary considerably in the degree to which they hear an /r/ on any occasion, a model based on a collection of acoustic cues can predict their group responses, and their confidence in judging /r/ presence, more accurately than any individual acoustic cue.

Second, the significant correlation between classifier probability and trained listeners' coding judgments further validates the use of a random forest classifier to perform automated coding of /r/. Together with the classifier's high rates of prediction accuracy (as demonstrated by cross-validation within training data), these experimental results provide validation of the method. This is methodologically welcome, given that the categorical coding of /r/ (and other sociolinguistic variables) is a time-consuming task that represents a bottleneck in the process of carrying out sociophonetic research. Further, while individual listeners appear to vary considerably in the degree to which they 'hear' /r/, probabilities from a model based on a single trained listener can capture group patterns accurately. This raises the intriguing possibility that a single trained listener's binary codes can be combined with the associated acoustics to generate gradient predictions which can accurately capture how additional trained listeners would code the tokens. If this is the case, this gradient metric would therefore be a more reliable measure of /r/ presence than an individual trained listener's binary coding decisions alone. In short, phonetically trained listeners hear /r/ as present to different degrees, and this variability in listener perceptions correlates well with a token's acoustic properties.[17]

[17] An editor rightly points out that we should exercise caution not to hastily generalize the findings of this experiment to what lay listeners are doing when they perceive /r/ tokens out in the world: "The relationship between what a trained expert decides when forced to make a binary classification and how listeners interpret the sociolinguistic content of the acoustic signal seems like it could be rather more complex than is being assumed here." These findings nevertheless indicate a promising degree of agreement between the classifier's predictions and a *particular* context of human perceptions of /r/, and the applicability to other contexts is an empirical question. To that end, we note conflicting findings between Hall-Lew & Fix (2012), who found little effect of coding experience on listeners' /l/-vocalization coding responses, and Yaeger-Dror et al. (2009), who did find a correlation between a linguist's exposure to /r/ variability and their /r/ coding responses.

6. General discussion and conclusion

In this paper, we have discussed the method of random forest classification as a means to automatically code sociophonetic variables, which is a welcome alternative to the tedious, time-consuming process of hand-coding. We have validated this method by showing that binary classifiers of two sociophonetic variables of English (non-prevocalic /r/ and word-medial intervocalic /t/) perform well in cross-validation and that the /r/ classifier's gradient predictions reflect variability in trained listeners' coding judgments. This method also sheds light on the low-level acoustic properties that best characterize variation in /r/ and /t/ in New Zealand English; contrary to the common practice of describing gradient variation in /r/ via F3 minimum, our /r/ classifier indicates that several measures involving differences between F3 and F1 or F2 at various timepoints are more informative in classification than F3 minimum. Over and above the importance of any individual acoustic measures, we find that classifier probability, as a composite of numerous acoustic measures, more accurately modeled trained listeners' coding judgments of /r/.

While it remains to be seen to what extent this finding of meaningful gradience in /r/ extends to other variables, this method raises the possibility of transforming a binary treatment of sociolinguistic variation to a continuous index of variability. This continuous treatment more accurately reflects the fact that individual tokens lie on an acoustic continuum that is not adequately captured by merely sorting tokens into Present versus Absent bins. As discussed in Section 3.1, there is growing evidence for treating /r/ as meaningfully gradient, focusing specifically on F3 minimum (Hashimoto, 2019; Hay & Clendon, 2012; Hay & Maclagan, 2010, 2012; Love & Walker, 2013). Moreover, as an editor points out, these findings questioning the assumption of categoricity in /r/ dovetail with recent evidence questioning the categoricity of English coronal stop deletion based on both acoustic (Temple, 2014) and articulatory (Purse, 2019) approaches. In light of this external evidence that gradience in English /r/ (and possibly other variables) is phonetically and sociolinguistically meaningful, the random forest classifier method represents a potential means to operationalize 'categorical' variation as gradient (at least where rhoticity is concerned) beyond individual acoustic correlates.

We further argue that treating 'categorical' sociophonetic variables as gradient surmounts another hurdle typically faced by hand-coding these variables as 'categorical': poor inter-rater reliability (see Section 3.3). Given a set of tokens, some listeners will assign larger or smaller numbers of tokens to one category or the other, in part due to asymmetries of expectation and experience among different listeners (Hay et al., 2018). In other words, while some tokens of a variable may clearly, unambiguously belong to a single class, for a considerable number of tokens there is no unambiguous ground truth. This is additional evidence in favor of reconceptualizing 'categorical' variation in probabilistic, continuous terms—and the listening experiment suggests that our classifier model successfully facilitates this reconceptualization. Out of 60 tokens, only one was categorically perceived as Absent by 11 listeners, and only one was categorically perceived as Present; the classifier was accurate in assigning these tokens very low and very high probabilities, respectively. Crucially, the tokens that listeners perceived in the gray area were assigned middling probabilities by the classifier. In a typical sociophonetic coding situation, tokens with such low levels of inter-coder agreement would be seen as problematic, requiring extensive intervention to help coders arrive at an agreed-upon 'correct' class; a continuous treatment, by contrast, recognizes the potential sociolinguistic meaningfulness of these gray area tokens. Lending weight to this interpretation is Hall-Lew and Fix's (2012) finding that the /l/ tokens that experienced the greatest disagreement among coders were those that were, on average, judged to be intermediate between vocalized and consonantal realizations. Moreover, it is worth remembering that the hand-coded classes that were used to train the classifier were generated by a single human rater (albeit a well-trained rater), but the probabilities generated by the model trained on this single rater's coding judgments nevertheless accurately predicted group patterns among a larger set of trained coders. These results suggest, as an empirical question, that this gradient metric may be a more reliable measure of rhoticity than any individual rater's binary ratings alone.

A gradient operationalization is not necessarily appropriate for all variables, however, as indicated by the differences in classifier probabilities for auto-coded /r/ versus /t/ tokens. While the distribution of classifier probabilities for auto-coded /r/ tokens was unimodal, with most tokens in the gray area between Absent and Present, auto-coded /t/ tokens' classifier probabilities were bimodally distributed, with most tokens clustered toward being highly Voiceless-like or highly Voiced-like. As a result, the probabilities returned by the /t/ classifier are less likely to meaningfully represent gradient variation between two endpoints than the /r/ classifier probabilities. In other words, while the /t/ classifier is well suited to generating binary predictions, the /r/ classifier's predictive capabilities are better suited to generating classifier probabilities. However, this does not eliminate the possibility that the /t/ classifier may be capturing interesting gradience *within* the two /t/ categories; some Voiced tokens, for example, are associated with a higher probability of voicing than others. It remains for further work to determine whether this variation is perceptually or linguistically meaningful.

Although we present what we find to be compelling proof-of-concept evidence for the use of random forest classifiers for automated coding of sociophonetic variation, we emphasize that readers should exercise caution when deploying this method on their own datasets.[18] A classifier's predictions are only as good as the coder(s) who create the training data, meaning that (as with traditional hand-coding) there is always an element of subjectivity involved. Even with unimpeachable training data, readers' mileage may vary with respect to classifier performance, partially as a function of the different properties of different variables. Indeed, while binary classifiers of /r/ and /t/ performed well, the /t/ classifier with six classes (some quite rare in the data) failed to perform up to acceptable levels, suggesting room for improvement in accounting for minority classes. Random forests' ability to handle collinear data, while useful for variables for which it is unclear which acoustic cues most meaningfully underlie variation, is also not an invitation to indiscriminately include every possible acoustic parameter in the model. In short, random forest classifiers are no 'magic wand.' As a result, we urge readers to conduct checks on the outputs of their classifiers (perhaps conducting their own listening experiments) before using this method to auto-code new tokens, assess variable importance, and/or generate gradient predictions.

Indeed, our intent with this paper is certainly *not* to have the last word on this method. While promising, our finding that the classifier method works for two variables (both of which have large training sets) in one variety is not an indication that this method is suitable for use in all conditions. Although a comparison of methods is outside the scope of this paper, machine-learning implementations of classifiers other than random forests may prove to be more appropriate for this application; indeed, as an editor suggests, a bottom-up clustering method may be more appropriate than a classification approach. It is further possible that, as anonymous reviewers suggest, classifiers trained on a smaller number of composite measures common in signal processing (e.g., discrete cosine transforms and mel-frequency cepstral coefficients) may outperform models trained on large numbers of 'traditional' acoustic-phonetic measures. Finally, while we intentionally

[18] Thanks to two anonymous reviewers for helping to clarify the ideas presented in this paragraph.

avoided introducing social or linguistic factors above the level of phonetics directly into our classifiers to prevent over-learning of top-down extra-phonetic features (see Section 4.2.1), it remains to be seen if extra-phonetic features can find their way indirectly into the model regardless (i.e., via the acoustic measures we did introduce into the model). As these are all rich areas for future study, our online classifier-training tutorial (https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html) contains the /r/ and /t/ data, and all the code needed to recreate the /r/ classifier reported in this paper, to allow researchers to explore and expand on this method on their own.

In conclusion, the method of random forest classification provides us with the ability to automatically code vast quantities of sociolinguistic variables with accuracy comparable to that of a trained linguist, opening research questions that would otherwise be time-prohibitive to pursue. To the extent that variables are meaningfully gradient (which remains an open question), this method also provides a means to capture that gradience more comprehensively than is possible in any individual acoustic measure. We argue that this method represents a potentially powerful new method in the sociolinguistic toolkit.

Acknowledgements

This research was made possible by a Royal Society of New Zealand Marsden Research Grant (16-UOC-058). The authors would like to thank members of the New Zealand Institute of Language, Brain and Behaviour for their feedback and support, particularly Viktória Papp for their assistance with random forests. Thanks are also due to Chris Bartlett and the Southland Oral History Project and Invercargill City Library for sharing data, as well as audiences at the Australian National University and the Variation and Language Processing 4 conference. This paper also benefited greatly from the thoughtful and incisive commentary of the *LabPhon* editors and two anonymous reviewers. Any errors are ours entirely.

Competing Interests

The authors have no competing interests to declare.

References

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology*, *8*(1). DOI: https://doi.org/10.5334/labphon.19

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bartlett, C. (2002). *The Southland Variety of New Zealand English: Postvocalic /r/ and the BATH vowel* (Thesis).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 48. DOI: https://doi.org/10.18637/jss.v067.i01

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38. DOI: https://doi.org/10.1016/j.wocn.2018.05.004

Becker, K. (2009). /r/ and the construction of place identity on New York City's Lower East Side. *Journal of Sociolinguistics*, *13*(5), 634–658. DOI: https://doi.org/10.1111/j.1467-9841.2009.00426.x

Boersma, P., & Weenink, D. (2015). *Praat*. Retrieved from http://www.fon.hum.uva.nl/praat/

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324

Brown, L., Winter, B., Idemaru, K., & Grawunder, S. (2014). Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*, *66*, 45–60. DOI: https://doi.org/10.1016/j.pragma.2014.02.011

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. DOI: https://doi.org/10.5334/joc.10

Buizza, E., & Plug, L. (2012). Lenition, fortition and the status of plosive affrication: The case of spontaneous RP English /t/. *Phonology*, *29*(1), 1–38. DOI: https://doi.org/10.1017/S0952675712000024

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*(1), 35. DOI: https://doi.org/10.1186/s13040-017-0155-3

Christensen, H. B. (2019). *Ordinal – Regression models for ordinal data*. Retrieved from http://www.cran.r-project.org/package = ordinal/

Clark, L. (2018). Priming as a motivating factor in sociophonetic variation and change. *Topics in Cognitive Science*, 1–16. DOI: https://doi.org/10.1111/tops.12338

Clark, L., MacGougan, H., Hay, J., & Walsh, L. (2016). "Kia ora. This is my earthquake story". Multiple applications of a sociolinguistic corpus. *Ampersand*, *3*, 13–20. DOI: https://doi.org/10.1016/j.amper.2016.01.001

Docherty, G., & Foulkes, P. (1999). Sociophonetic variation in 'glottals' in Newcastle English. *14th ICPhS*, 1037–1040. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_1037.pdf

Dodsworth, R., & Kohn, M. (2012). Urban rejection of the vernacular: The SVS undone. *Language Variation and Change*, *24*(2), 221–245. DOI: https://doi.org/10.1017/S0954394512000105

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P., Münkemüller, T., McClean, C., Osborne, P., Reineking, B., Schröder, B., Skidmore, A., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. DOI: https://doi.org/10.1111/j.1600-0587.2012.07348.x

Fiasson, R. (2015). *Allophonic imitation within and across word positions* (Thesis). Retrieved from https://ir.canterbury.ac.nz/handle/10092/11514

Fosler-Lussier, E., Dilley, L., Tyson, N. R., & Pitt, M. A. (2007). The Buckeye Corpus of Speech: Updates and enhancements. *Interspeech*, *8*, 934–937. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2007/i07_0934.pdf

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Retrieved from http://tinyurl.com/carbook

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012). Automated sleep stage identification system based on time – frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10–19. DOI: https://doi.org/10.1016/j.cmpb.2011.11.005

Fromont, R., & Hay, J. (2012). LaBB-CAT: An annotation store. *Proceedings of Australasian Language Technology Association Workshop*, 113–117. Retrieved from http://www.aclweb.org/anthology/U12-1015

German, J. S., Carlson, K., & Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics*, *41*(3), 228–248. DOI: https://doi.org/10.1016/j.wocn.2013.03.001

Gibson, A. (2005). Non-prevocalic /r/ in New Zealand hip hop. *New Zealand English Journal*, *19*, 5–12.

Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., & Trudgill, P. (2004). *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511486678

Hall-Lew, L., & Fix, S. (2012). Perceptual coding reliability of (L)-vocalization in casual speech data. *Lingua*, *122*(7), 794–809. DOI: https://doi.org/10.1016/j.lingua.2011.12.005

Harrell, F. E. (2018). *Rms: Regression Modeling Strategies*. Retrieved from https://CRAN.R-project.org/package = rms

Hashimoto, D. (2019). *Loanword phonology in New Zealand English: Exemplar activation and message predictability* (Thesis). Retrieved from https://ir.canterbury.ac.nz/handle/10092/16634

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Retrieved from https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Hay, J., & Clendon, A. (2012). (Non)rhoticity: Lessons from New Zealand English. In T. Nevalainen & E. C. Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 761–772). Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199922765.013.0063

Hay, J., Drager, K., & Gibson, A. (2018). Hearing r-sandhi: The role of past experience. *Language*, *94*(2), 360–404. DOI: https://doi.org/10.1353/lan.2018.0020

Hay, J., & Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language*, *92*(2), 298–330. DOI: https://doi.org/10.1353/lan.2016.0036

Hay, J., & Maclagan, M. (2010). Social and phonetic conditioners on the frequency and degree of 'intrusive /r/' in New Zealand English. In D. R. Preston & N. A. Niedzielski (Eds.), *A reader in sociophonetics* (pp. 41–69). New York: De Gruyter Mouton.

Hay, J., & Maclagan, M. (2012). /r/-sandhi in early 20th century New Zealand English. *Linguistics*, *50*(4), 745–763. DOI: https://doi.org/10.1515/ling-2012-0023

Hay, J., & Sudbury, A. (2005). How rhoticity became /r/-sandhi. *Language*, *81*(4), 799–823. Retrieved from http://www.jstor.org/stable/4490019. DOI: https://doi.org/10.1353/lan.2005.0175

Heselwood, B. (2009). Rhoticity without F3: Lowpass filtering and the perception of rhoticity in 'NORTH/FORCE,' 'START,' and 'NURSE' words. *Leeds Working Papers in Linguistics and Phonetics*, *14*, 49–64.

Holmes, J. (1994). New Zealand flappers: An analysis of T voicing in New Zealand English. *English World-Wide*, *15*(2), 195–224. DOI: https://doi.org/10.1075/eww.15.2.03hol

Irwin, R. B. (1970). Consistency of judgments of articulatory productions. *Journal of Speech and Hearing Research*, *13*(3), 548–555. DOI: https://doi.org/10.1044/jshr.1303.548

Jones, M. J., & Llamas, C. (2008). Fricated realisations of /t/ in Dublin and Middlesbrough English: An acoustic analysis of plosive frication and surface fricative contrasts. *English Language and Linguistics*, *12*(3), 419–443. DOI: https://doi.org/10.1017/S1360674308002700

Jones, M. J., & McDougall, K. (2009). The acoustic character of fricated /t/ in Australian English: A comparison with /s/ and /ʃ/. *Journal of the International Phonetic Association*, *39*(3), 265–289. DOI: https://doi.org/10.1017/S0025100309990132

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625. DOI: https://doi.org/10.1146/annurev-psych-122414-033702

Kennedy, M. (2006). *Variation in the pronunciation of English by New Zealand school children* (Thesis). Retrieved from https://core.ac.uk/download/pdf/41335595.pdf

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics, 70*, 70–85. DOI: https://doi.org/10.1016/j.wocn.2018.05.005

Kuhn, M. (2018). *Caret*. Retrieved from https://CRAN.R-project.org/package=caret

Kuznetsova, A., Brockhoff, B., & Christensen, H. B. (2016). *lmerTest*. Retrieved from https://CRAN.R-project.org/package=lmerTest

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110167467

Labov, W., Rosenfelder, I., & Freuhwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65. DOI: https://doi.org/10.1353/lan.2013.0015

Lawson, E., Scobbie, J., & Stuart-Smith, J. (2014). A socio-articulatory study of Scottish rhoticity. In R. Lawson (Ed.), *Sociolinguistics in Scotland* (pp. 53–78). London: Palgrave Macmillan. DOI: https://doi.org/10.1057/9781137034717_4

Lawson, E., Stuart-Smith, J., & Scobbie, J. (2018). The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *Journal of the Acoustical Society of America*, *143*(3), 1646–1657. DOI: https://doi.org/10.1121/1.5027833

Love, J., & Walker, A. (2013). Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech*, *56*(4), 443–460. DOI: https://doi.org/10.1177/0023830912453132

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33. DOI: https://doi.org/10.1080/10888438.2015.1107073

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *18th Conference of the International Speech Communication Association. Conference Proceedings*. DOI: https://doi.org/10.21437/Interspeech.2017-1386

McLarty, J., Jones, T., & Hall, C. (2019). Corpus-based sociophonetic approaches to postvocalic r-lessness in African American Language. *American Speech*, *94*. DOI: https://doi.org/10.1215/00031283-7362239

Nagy, N., & Irwin, P. (2010). Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change*, *22*(2), 241–278. DOI: https://doi.org/10.1017/S0954394510000062

Nielson, D., & Hay, J. (2005). Perceptions of regional dialects in New Zealand. *Te Reo*, *48*, 95–110.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690. DOI: https://doi.org/10.1007/s11135-006-9018-6

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye Corpus of Conversational Speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. DOI: https://doi.org/10.1016/j.specom.2004.09.001

Purse, R. (2019). The articulatory reality of coronal stop "deletion". In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), *19th International Congress of Phonetic Sciences* (pp. 1595–1599). Retrieved from https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1644.pdf

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/. DOI: https://doi.org/10.3115/v1/N15-3015

Reddy, S., & Stanford, J. (2015). A web application for automated dialect analysis. In *Proceedings of NAACL-HLT 2015*. Retrieved from http://www.aclweb.org/anthology/N15-3015

Riehl, A. K. (2003). American English flapping: Evidence against paradigm uniformity with phonetic features. *15th ICPhS*, 2753–2756. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2753.pdf

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. DOI: https://doi.org/10.1016/j.isprsjprs.2011.11.002

Rosenberg, A. (2017). *AuToBI: Automatic prosodic annotation*. Retrieved from https://github.com/AndrewRosenberg/AuToBI

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) program suite*. Retrieved from http://fave.ling.upenn.edu/

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. DOI: https://doi.org/10.2307/3002019

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96–109. DOI: https://doi.org/10.1016/j.wocn.2010.11.006

Schuppler, B., van Dommelen, W. A., Koreman, J., & Ernestus, M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*(4), 595–607. DOI: https://doi.org/10.1016/j.wocn.2012.05.004

Seyfarth, S., & Garellek, M. (2015). Coda glottalization in American English. *18th ICPhS*. Conference Proceedings. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0807.pdf

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *6th International Conference on Language Resources and Evaluation*. Conference Proceedings. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf

Sonderegger, M., & Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured prediction. *The Journal of the Acoustical Society of America*, *132*(6), 3965–3979. DOI: https://doi.org/10.1121/1.4763995

Sonderegger, M., Wagner, M., & Torreira, F. (2018). *Quantitative methods for linguistic data*. Retrieved from http://people.linguistics.mcgill.ca/~morgan/book/index.html

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307. DOI: https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. (Supplemental). DOI: https://doi.org/10.1037/a0016973

Strobl, C., & Zeileis, A. (2008). Danger: High power! Exploring the statistical properties of a test for random forest variable importance. *18th International Conference on Computational Statistics*. Conference Proceedings. Retrieved from https://epub.ub.uni-muenchen.de/2111/

Stuart-Smith, J. (2007). A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. In J. Trouvain & W. J. Barry (Eds.), *16th ICPhS* (pp. 1449–1452).

Stuart-Smith, J., Lawson, E., & Scobbie, J. (2014). Derhoticisation in Scottish English: A sociophonetic journey. In C. Celata & S. Calamai (Eds.), *Advances in sociophonetics* (pp. 59–96). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/silv.15.03stu

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178. DOI: https://doi.org/10.1017/S0954394512000129

Temple, R. A. M. (2014). Where and what is (t, d)? A case study in taking a step back in order to advance sociophonetics. In *Advances in Sociophonetics* (pp. 97–136). Retrieved from http://ebookcentral.proquest.com/lib/canterbury/detail.action?docID = 1715253. DOI: https://doi.org/10.1075/silv.15.04tem

van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics*, *32*(4), 455–491. DOI: https://doi.org/10.1016/j.wocn.2004.05.001

Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustical Society of America*, *130*(3), 1606–1617. DOI: https://doi.org/10.1121/1.3621306

Wei, H., Cheong-Fat, C., Chiu-Sing, C., & Kong-Pang, P. (2006). An efficient MFCC extraction method in speech recognition. *2006 IEEE International Symposium on Circuits and Systems*, 4 pp. DOI: https://doi.org/10.1109/ISCAS.2006.1692543

Wells, J. C. (1982). *Accents of English*. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511611759

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. DOI: https://doi.org/10.1037/xge0000014

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C+ + and R. *Journal of Statistical Software*, *77*(1), 1–17. DOI: https://doi.org/10.18637/jss.v077.i01

Yaeger-Dror, M., Kendall, T., Foulkes, P., Watt, D., Oddie, J., Johnson, D. E., & Harrison, P. (2009). *Perception of 'r': A cross-dialect comparison*.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2006). *The HTK book (for HTK version 3.4)*. Cambridge University Engineering Department.

Zamani, Y., Souri, Y., Rashidi, H., & Kasaei, S. (2015). Persian handwritten digit recognition by random forest and convolutional neural networks. *2015 9th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 37–40. DOI: https://doi.org/10.1109/IranianMVIP.2015.7397499

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *Journal of the Acoustical Society of America*, *123*(6), 4466–4481. DOI: https://doi.org/10.1121/1.2902168

How to cite this article: Villarreal, D., Clark, L., Hay, J., & Watson, K. 2020 From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 6, pp. 1–31. DOI: https://doi.org/10.5334/labphon.216

Submitted: 24 July 2019 **Accepted:** 11 April 2020 **Published:** 10 June 2020

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS