MODELLING GRADIENCE IN ENGLISH /r/ VIA STATISTICAL CLASSIFICATION

Dan Villarreal, Lynn Clark, Jennifer Hay

University of Canterbury
daniel.villarreal@canterbury.ac.nz; lynn.clark@canterbury.ac.nz; jen.hay@canterbury.ac.nz

ABSTRACT

This experiment assessed the validity of a statistical classification method for automated coding of sociophonetic variables, in particular the presence vs. absence of English non-prevocalic /r/. A random forest classifier was trained on 180 acoustic measures from 5,355 tokens of /r/ (hand-coded as Present or Absent) in a variety of English with variable rhoticity; the classifier achieved 87.9% accuracy on training data. The classifier was then used to predict the probability of /r/ presence in 32,099 additional tokens from the same variety.

Eleven phonetically trained listeners judged 60 classifier-coded tokens as Present or Absent. Judgment results indicated a significant positive linear relationship between classifier probability and human judgments; classifier probability also outperformed individual acoustic measures (e.g., F3 minimum) in predicting human judgments. These results both validate this random forest classifier method for automated coding of sociophonetic variables and indicate the viability of modelling phonetic variation using classifier probability.

Keywords: rhoticity, random forests, speech perception, sociophonetics, acoustic phonetics

1. BACKGROUND

English non-prevocalic /r/ is acoustically complex, but the typical approach to understanding variable rhoticity in sociolinguistics is to treat it as binary, varying between *Present* and *Absent* (or rhotic/*r*-ful and non-rhotic/*r*-less) [1, 3, 6, 12, 17]. Under this approach, inter- and intraspeaker variation in /r/ is represented in terms of proportions, with speakers and varieties said to be more rhotic if a greater proportion of their /r/ tokens are Present. However, inter-coder reliability is notoriously low for rhoticity, even among phoneticians [13, 27].

In cases where sociolinguistic research treats /r/ as continuous, it is typically in terms of F3 minimum, with lower F3 corresponding to a greater constriction and thus a 'stronger' /r/ [8, 15]. By contrast, approaches to English /r/ in acoustic phonetics have found a complex array of cues to rhoticity: a greater lag between F3 minimum and the end of the token [14], longer duration and lower F2 [22], spectral information below F3 [9], and a lack of frication noise immediately after the token [23]. Complicating matters further, rhoticity subsumes distinct articulatory strategies with different acoustic consequences [13, 28]. In sum, despite widespread intuition that /r/ is not a categorical variable, it is not at all clear that F3 is the best means for modelling continuous variation in /r/. Indeed, it is likely that the acoustic complexity and heterogeneity of /r/ is in part what makes /r/ so difficult to code reliably.

We trained a random forest classifier on tokens of variable rhoticity in natural speech that were coded as /r/ Present or Absent. The classifier then used 180 acoustic measurements contained in the training set to decide whether uncoded /r/s in natural speech should be coded as Present or Absent; the model produces, for each token, a probability that it contains /r/ (*classifier probability*). Tokens that are assigned high probabilities by the classifier strongly resemble, in multiple ways, tokens of /r/ known to be Present. Tokens with lesser probabilities contain fewer properties that are typical of the acoustic signature of /r/. Our initial goal was to efficiently produce binary codes for a large dataset, allowing us to proceed more quickly with a robust sociolinguistic analysis. This paper explores the additional possibility that the random forest model can also be used as a composite measure of the gradient presence of /r/. We test the hypothesis that, in addition to providing a binary categorization for uncoded data, classifier probability may be a useful and meaningful index of gradient rhoticity in its own right. Our research questions are:

1. What is the relationship between classifier probabilities and human judgments of English non-prevocalic /r/?
2. How does classifier probability compare to individual acoustic measures of /r/ in predicting human judgments of /r/?

In order to explore these questions, we selected 60 classifier-coded tokens for a perceptual task in which 11 phonetically trained listeners judged tokens as Present or Absent and rated their confidence. We performed statistical modelling of these judgments and confidence ratings to determine to what extent classifier probability and individual acoustic measures predicted human responses to /r/; we find significant agreement between classifier probability and human judgments. We thus see this method as addressing both a practical need (automated

annotation of large data sets) and a theoretical need (determining which acoustic properties best characterise /r/).

2. RANDOM FOREST CLASSIFIER

Previous research has found variable rhoticity in the Southland region of New Zealand [1]. Our corpus of Southland English includes over 83 hours of sociolinguistic and oral history interviews from speakers born 1868-1998. It includes several thousand /r/ tokens that were hand-coded by a single researcher. We fit a random forest classifier to these hand-coded tokens.

Random forests are an extension of classification and regression trees, which recursively partition data into successively smaller subsets at each tree node by finding the independent variable that minimises variation in the subbranches under the node. Individual trees select from among random subsets of independent variables at each split. The ensembled trees in a forest are averaged to a consensus on which predictors are most important [24]. Unlike many modelling techniques, random forests do not suffer when predictors are collinear [16, 21]. Random forest models can also perform classification on unseen data [see e.g. 19].

Our random forest classifier was trained on 5,355 hand-coded tokens of non-prevocalic /r/ from 28 individual speakers (89–229 tokens per speaker; 28.3% of all tokens coded as present). This training set excluded tokens for which there were missing measurements, or for which F3 measurements at the 25%, 35%, 75%, or 80% timepoints were outliers.

For each token, we extracted 180 acoustic measures across the sequence of vowel plus possible /r/ (e.g., *st<u>ar</u>t*, *n<u>ur</u>se*). Our choice of features was guided by two aims: producing a well-performing classifier, and resolving persistent uncertainty about the acoustic features that best characterise /r/ [e.g., 8, 9, 14, 22, 23, 28]. The latter aim also meant that we did not introduce any social factors or linguistic factors above the level of phonetics (e.g., speaker gender, vowel phoneme, stress) into the classifier; we also wanted to avoid the classifier over-learning extra-phonetic features, as these features may have applied in different degrees to the training vs. test sets. This second aim is also why we did not use features like mel-frequency cepstral coefficients that are popular in the fields of signal processing and speech recognition [e.g., 25], opting instead for measures that have more currency in acoustic phonetics. The measures were:

- Formant measurements at 13 timepoints. These measurements were normalised by subtracting the speaker's mean word-initial /r/ midpoint measurement for that formant from the raw measurement.
- Formant maxima and minima (speaker normalised) and the normalised time at which maxima and minima were found
- Formant ranges (raw maximum minus raw minimum), and slopes (range divided by normalised time)
- Differences between raw formant values at 13 timepoints (e.g., differences between F2 and F1, F3 and F1)
- Formant bandwidths at 13 timepoints
- Pitch maxima and minima, normalised timepoints of maxima and minima, pitch range, and pitch slope
- Amplitudes at F3 maxima and minima
- Token duration, *z*-scored by speaker.

These measures were entered into a random forest in R using the packages ranger and caret [10, 18, 26]. Binary /r/ was the dependent variable. This forest included 1,000 trees, tested 13 variables at each node, used a Gini splitting rule, had no minimum node size restriction, and measured variable importance via the Gini index.

To test classifier performance against hand-codes, classifier probabilities were converted to binary codes by specifying a probability cutoff at which tokens would be coded Present. Preliminary testing found that accuracy was optimised by using a probability cutoff at 0.579065. Within the training set, prediction accuracy was assessed via cross-validation; in four rounds, the training set was split into training and test subsets via the 0.632 bootstrap estimator and SMOTE sampling to resolve the imbalance in Present and Absent classes [4, 5]. This procedure found a prediction accuracy of 87.9%, which compares favourably with human coders' 84–86% inter-rater reliability for /r/ [13].

We then used the classifier to predict 32,099 additional tokens of non-prevocalic /r/ in our corpus. The classifier coded 19.8% of these tokens as Present, with a mean classifier probability of 0.40.

3. JUDGMENT TASK

3.1. Stimuli and task

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. We 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. Sixty such tokens were selected to span the range of classifier

probabilities, with 32 tokens coded Absent and 28 coded Present by the classifier. Stimuli were created by extracting the relevant word from the audio file, resampling the word to 22050 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 750ms buffer between repetitions. Listeners performed the task with headphones on individual computers. All listeners were proficient English users who lived in NZ at the time of the experiment. Five listeners self-reported speaking English with rhotic accents, six with non-rhotic accents; four listeners self-reported as non-native speakers.

3.2. Analysis

Listeners' proportion of Present judgments was highly variable (M = 0.5038, SD = 0.1897, range = 0.13–0.87). Mixed-effects modelling was performed in R [18]; judgments data (binary) were modelled via logistic regression in the lme4 package [2], and confidence data (continuous) via linear regression with Satterthwaite approximations for degrees of freedom in the lmerTest package [11, 20]. 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; additionally for confidence, the stimulus random effect was nested within word and there was a random slope of classifier probability by listener. Confidence was modelled separately for stimuli judged Present (n = 332) vs. Absent (n = 327).

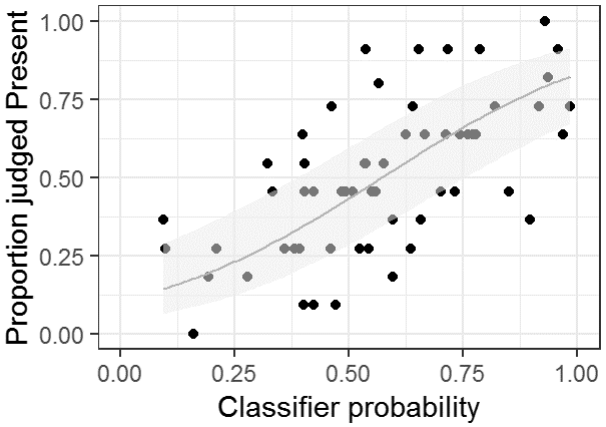
We also modelled classifier probability as a restricted cubic spline with three knots in the R package rms [7], but these were never significantly better than the baseline models. Finally, to address research question 2, we ran models of judgments where classifier probability was replaced by individual acoustic measures; since these models were not nested within the baseline model, their log-likelihoods were compared to that of the baseline.

4. RESULTS

4.1. Stimuli and task

The baseline model of judgments revealed a significant positive effect of classifier probability on human judgments (β = 3.73, z = 5.91, p < .0001), indicating that stimuli with greater classifier probabilities were more likely to be judged Present; this relationship is evident in Figure 1 below. Following segment was not significant (ps > .60). 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 indicating a linear rather than nonlinear relationship ($\chi^2[1] = 0.02$, p = 0.89).

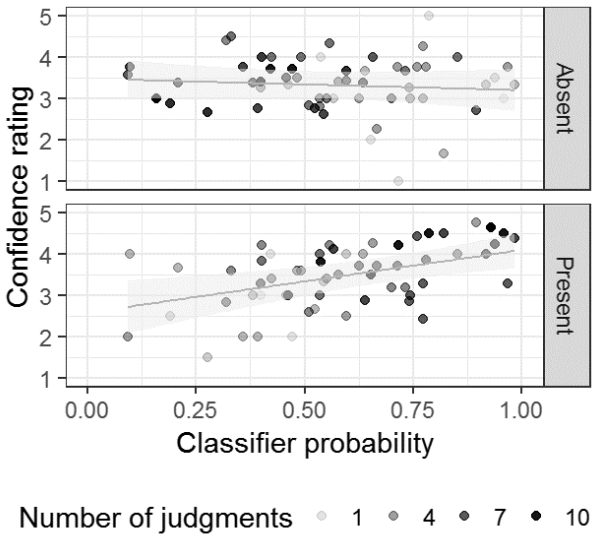
Figure 1: Classifier probability vs. human judgments for each stimulus (dots), with fitted effects line and 95% confidence band from best model of judgments.



4.2. Confidence

The baseline model of confidence for stimuli judged Present revealed a significant positive effect of classifier probability on confidence ratings (β = 1.51, $t[14.63] = 3.14$, p < .01), indicating that listeners were more confident in judging stimuli with greater classifier probabilities as Present. However, in the baseline model of confidence for stimuli judged Absent, classifier probability had no significant effect on confidence ratings (β = −0.28, $t[9.60] = -0.68$, p = .51), indicating that listeners' confidence when judging stimuli as Absent was not related to classifier probability. This relationship is evident in Figure 2.

Figure 2: Classifier probability vs. mean confidence rating for stimuli judged Absent vs. Present (darker dots received more judgments), with fitted effects lines and 95% confidence bands.



4.3. Individual measures

To test whether classifier probability improves on individual acoustic measures, we ran three additional models of judgments, where classifier probability was replaced by (*z*-scored) individual acoustic measures: raw and speaker-normalized F3 minimum; and F3-F1 at 60% of the token's duration (the top-ranked variable in importance in the classifier). In all of these models, the individual measure significantly affected human judgments in the expected (negative) direction (*p*s < .005). However, the model with classifier probability as the main predictor outperformed them all (classifier log-likelihood: −383.72, individual measures: −387.8–−394.39).

Of course, 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 measurements. However, repeating the above comparisons with subsamples to match for distribution shapes across different predictors still points to the superiority of the classifier model (classifier log-likelihood: −151.56, individual measures: −160.50–−158.91).

5. DISCUSSION

Predictions from a random forest classifier trained on binary coded data can accurately predict gradient responses of human listeners.

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 behaviour as a group (cf. figure 1). And second, when individual listeners did hear an /r/ as Present, the classifier was able to predict how confident they were in that judgement. This 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 individual acoustic cues that best predict humans' binary classification of /r/. The difference between F3 and F1 shortly after the midpoint emerged as the cue that contributed to classifier performance most. However, listener behaviour was best predicted not by any individual acoustic cue, but by the classifier probability, which was able to consider a range of acoustic properties. 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. While our human 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 human 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 listener can capture group patterns accurately. A single listener's binary ratings can therefore be combined with the associated acoustics to generate gradient predictions which can accurately capture how a wider community would perceive the tokens. This gradient metric therefore seems likely to be a much more reliable measure of /r/ presence than the individual rater's binary ratings alone

Finally, the fact that classifier probability and human judgments exhibit a *linear* relationship adds weight to the interpretation of /r/ as an inherently gradient, rather than binary sociolinguistic variable. Listeners hear /r/ as present to different degrees, and this variability correlates well with a token's acoustic properties.

6. REFERENCES

[1] Bartlett, C. 2002. The Southland Variety of New Zealand English: Postvocalic /r/ and the BATH vowel. Unpublished PhD thesis. University of Otago.

[2] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.* 67, 48.

[3] Becker, K. 2009. /r/ and the construction of place identity on New York City's Lower East Side. *J. Socioling.* 13, 634-658.

[4] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 321-357.

[5] Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* 78, 316-331.

[6] Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., Trudgill, P. 2004. *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press.

[7] Harrell, F.E. 2018. rms: Regression Modeling Strategies, vers. 5.1-2. https://CRAN.R-project.org/package=rms

[8] Hay, J., Maclagan, M. 2012. /r/-sandhi in early 20th century New Zealand English. *Linguistics* 50, 745-763.

[9] 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.

[10] Kuhn, M. 2018. caret, vers. 6.0-81. https://CRAN.R-project.org/package=caret

[11] Kuznetsova, A., Brockhoff, B., Christensen, H.B. 2016. lmerTest, vers. 2.0-33. https://CRAN.R-project.org/package=lmerTest

[12] Labov, W., Ash, S., Boberg, C. 2006. *The Atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.

[13] Lawson, E., Scobbie, J., Stuart-Smith, J. 2014. A socio-articulatory study of Scottish rhoticity. In: R. Lawson, (ed), *Sociolinguistics in Scotland*. London: Palgrave Macmillan, 53-78.

[14] Lawson, E., Stuart-Smith, J., Scobbie, J. 2018. The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *J. Acoust. Soc. Am.* 143, 1646-1657.

[15] Love, J., Walker, A. 2013. Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech* 56, 443-460.

[16] 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, 20-33.

[17] Nagy, N., Irwin, P. 2010. Boston (r): Neighbo(r)s nea(r) and fa(r). *Lang. Var. and Change* 22, 241-278.

[18] R Core Team. 2018. R: A language and environment for statistical computing, vers. 3.5.1. https://www.R-project.org/

[19] 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.

[20] Satterthwaite, F.E. 1946. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* 2, 110-114.

[21] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307.

[22] Stuart-Smith, J. 2007. A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. *Proc. 16th ICPhS* Saarbrücken, Germany, 1449-1452.

[23] Stuart-Smith, J., Lawson, E., Scobbie, J. 2014. Derhoticisation in Scottish English: a sociophonetic journey. In: C. Celata and S. Calamai, (eds), *Advances in sociophonetics*. Amsterdam: John Benjamins, 59-96.

[24] Tagliamonte, S.A., Baayen, R.H. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Lang. Var. and Change* 24, 135-178.

[25] Wei, H., Cheong-Fat, C., Chiu-Sing, C., Kong-Pang, P. 2006. An efficient MFCC extraction method in speech recognition. *Proc. 2006 IEEE International Symposium on Circuits and Systems*, 4 pp.

[26] Wright, M.N., Ziegler, A. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Soft.* 77, 1-17.

[27] Yaeger-Dror, M., Kendall, T., Foulkes, P., Watt, D., Oddie, J., Johnson, D.E., Harrison, P. 2009. Perception of 'r': A cross-dialect comparison. *Linguistic Society of America* San Francisco.

[28] 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/. *J. Acoust. Soc. Am.* 123, 4466-4481.