Sociophonetics Autocoding

Yunting Gu, Shuangyu Zhao, Wenting Liu

1.Introduction

In sociophonetics studies – the research that concerns the sound variation of languages, it is almost always necessary to code the linguistic variables – the sound that may vary by speakers. However, hand-coding the sociolinguistic variables is timing consuming and could be inaccurate since the human coder's decision may be affected by their prior linguistic experience. This projects aims to test models that can auto-code the linguistic variables.

One target sound for auto-coding is the /r/ sound or rhotacization– it relates to the letter *r* in words in many cases.[1] For instance, in the word *water*, people may pronounce it with r as in *water* or without the rhotic as in *wate*. The /r/ sound is acoustically complex and can serve as a sociolingusitic variable crosslinguistically. In New Zealand English (NZE) (and numerous Englishes worldwide), /r/ is sociolinguistically meaningful, meaning that the usage of /r/ correlates with the social class, gender, age, or other social identity of the speakers. One purpose of the project is to test different models' behavior on classifying the present versus absent of /r/.

On the other hand, the advancement of speech recognition technology can also aid sociophonetic studies and reduce workload of manual coding. The second purpose of the project is to train a classifier to identify whether a given sound is /r/ or /t/ in New Zealand English. The finding of the study could aid future sociolinguistic research and inform acoustic studies about how to efficiently distinguish the sounds.

2.Dataset Description and Research Questions

2.1 Dataset Description

There are two raw datasets that come from the sociolinguistic literature Villarreal et al. (2020): one dataset contains the target variable /r/, and another dataset has the information for the variable /t/. Dataset 1 for the /r/ variable has 40,614 rows and 217 columns, and most of the columns contain acoustic features which are numerical values. It has 4,255 rows which are hardcoded as the absent of /r/, as well as 1,646 rows hardcoded as the present of /r/. Other rows in dataset 1 has not yet been coded for absent or present. Dataset 2 has information for

[1] Note that the slash line means that the bracketed letter is a sound symbol.

the /t/ variable. It has 9,888 rows and 137 column, and most columns have numerical values which are acoustic features.

2.2 Research Questions

There are two questions that is the focus of ths study. Question 1 is: What is the best model to classify the absence or presence of the sociolinguistic variable /r/? Question 2 is: What is the best model to distinguish between the /r/ sound and the /t/ sound?

We aim to answer two research questions above using the New Zealand English datasets. Since /r/ could serve as sociolinguistic variable in many varieties of English and also in other languages, figuring out what is the best model for question 1 could help researchers who are interested the /r/ variable in other languages as well. Question 2 contributes to the understanding of speech recognition.

3.Data Preprocessing and EDA

3.1 Data Preprocessing

Since two questions are addressed in this project, two new datasets need to be created based on the original ones. For question 1, our focus is on the dataset related to the information on /r/. As specified in the question description, the goal is to determine the presence or absence of /r/. The target variable is encoded as "absent" = 0 and "present" = 1. The rows not coded for presence or absence in the original dataset are removed. As for the columns, based on our understanding of the original dataframe, only the numeric features, which are acoustic information extracted from recordings, are the most relevant and worth encoding. Therefore, only the numeric features are retained, and they are standardized and normalized in the end. The dataset for question 1 has 187 columns and 5,273 rows.

For question 2, two datasets - one for /r/ and the other for /t/ - need to be merged. To achieve this, the column names of both datasets are converted to lowercase, and then they are merged based on matching column names. A target column is then created, with /r/ assigned the value 1 and /t/ assigned the value 0. Finally, only the numeric features are retained, and they are standardized and normalized. The dataset for question 2 has 52 columns and 49,753 rows.

3.2 Exploratory Data Analysis

3.2.1 The Dataset for Question 1

The target for Question 1 is to determine the presence or absence of /r/. Figure 1 displays the distribution of these two classes, with a present-to-absent ratio of 1:2, indicating an acceptable balance.

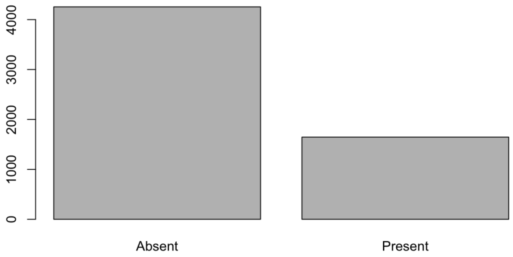
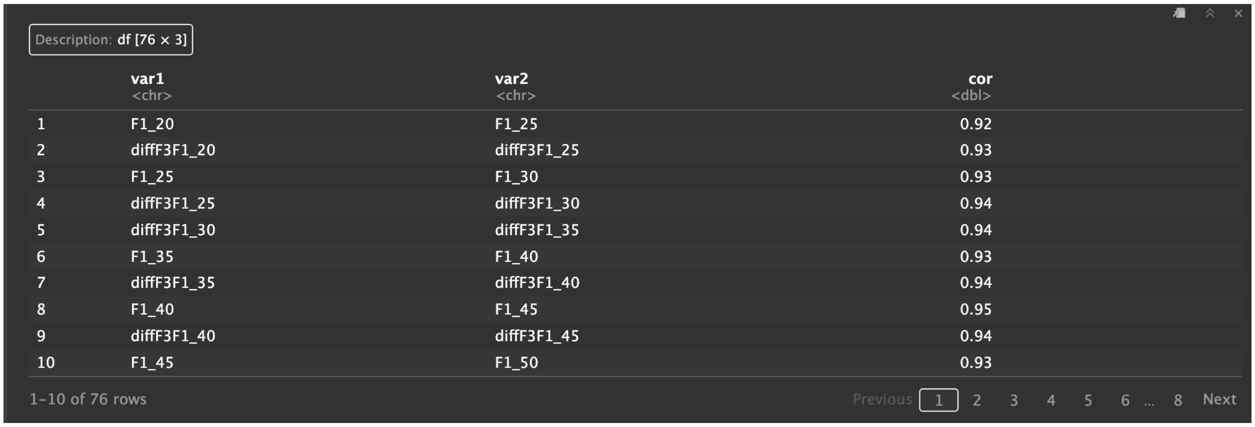


Figure 1 the barplot about the numbers of two classes in dataset for question 1

This dataset comprises 184 features, and their correlations are calculated. Figure 2 illustrates that 76 pairs of features have correlations greater than 0.9. This is due to the fact that the numeric features in the dataset are extractions of acoustic signals at 5% increment points, and adjacent extractions are highly correlated. For instance, in the first row of Figure 2, F1_20 (i.e., F1 or first formant value at 20% point) and F1_25 (F1 value at 25% point) are highly correlated. Figure 3a displays the correlation between the features and the target, with all correlations below 0.5. Figure 3b illustrates the distribution of intens_F3min, which has the highest correlation with the target. F3min means the lowest point of the third formant, which is considered as the acoustic correlation of /r/ or presence of /r/, so it is not surprising that it is found to be the highest correlated feature.



Description: df [76 × 3]

	var1	var2	cor
	<chr>	<chr>	<dbl>
1	F1_20	F1_25	0.92
2	diffF3F1_20	diffF3F1_25	0.93
3	F1_25	F1_30	0.93
4	diffF3F1_25	diffF3F1_30	0.94
5	diffF3F1_30	diffF3F1_35	0.94
6	F1_35	F1_40	0.93
7	diffF3F1_35	diffF3F1_40	0.94
8	F1_40	F1_45	0.95
9	diffF3F1_40	diffF3F1_45	0.94
10	F1_45	F1_50	0.93

1–10 of 76 rows

Figure 2 couples of features whose correlations are highly correlated in dataset for question 1

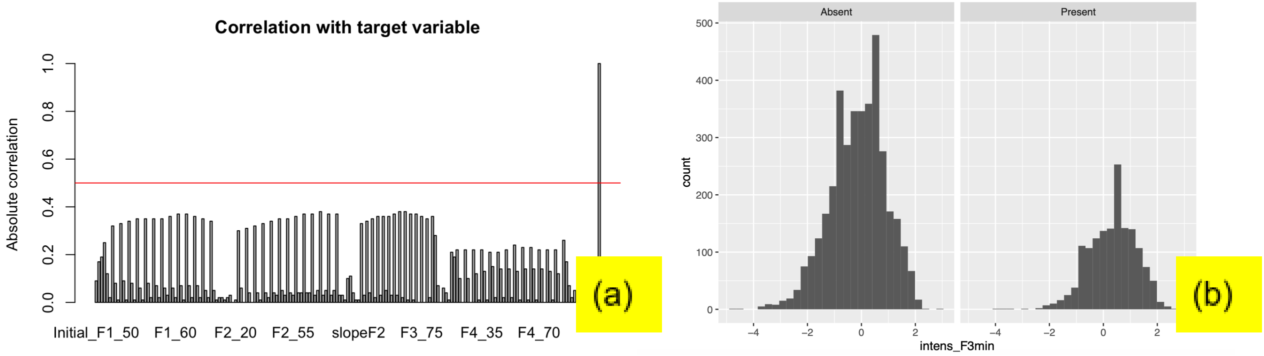


Figure 3 mappings about target correlation
(a. Correlations between features and target; b. Distribution of the feature who has the highest correlation with target)

3.2.2 The Dataset for Question 2

Figure 4 displays the distribution of two classes of target column, with a /t/-to-/r/ ratio of 1:4, which indicates the high imbalance of two classes. Therefore, in the model estimation part, different assessment indicators are chosen for question 2.

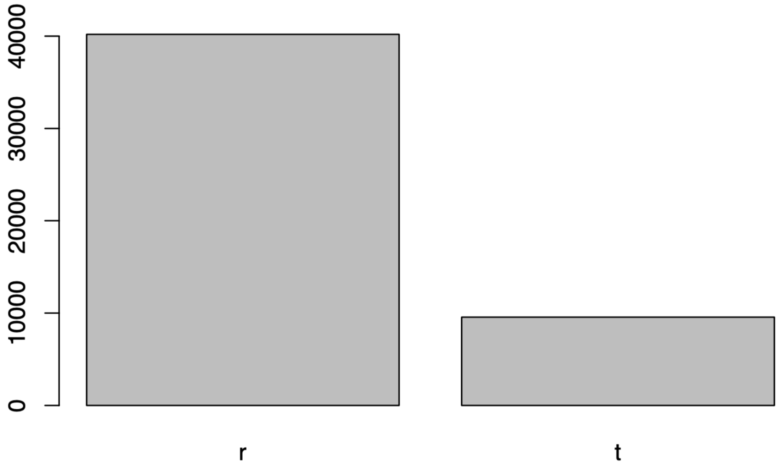
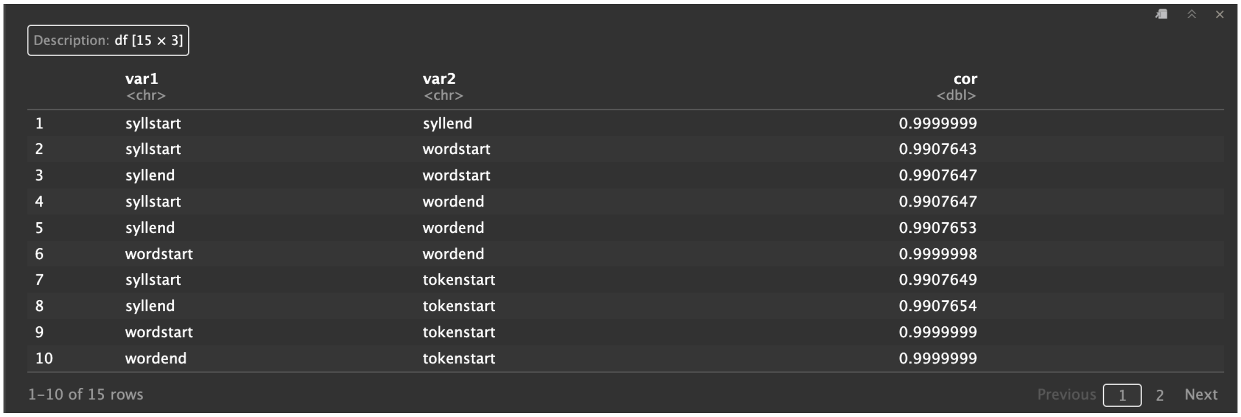


Figure 4 the barplot about the numbers of two classes in dataset for question 2

This dataset has 49 features. Figure 5 shows that 15 couples of features correlations are higher than 0.9. In Figure 5, syll is short for syllable, so the correlated features are likely not acoustic features.[2] Rather, they are timestamp information that indicates the starting points or end points of syllable, so the starting timestamp of a syllable (i.e., syllstart) is highly correlated with the end point of a syllable (i.e., syllend). Figure 6 displays the correlation between the features and the target, with two correlations higher than 0.5 and the others below 0.5. The two plots in Figure 7 are the distributions of two features whose correlations with target are higher than 0.5. The

[2] A vowel is roughly a syllable. In the word badminton, there are 3 syllables bad-min-ton.

feature tokennum is not very meaningful acoustically because the correlation can come from the fact that two datasets have different ways of numbering its token. The feature of celexfreq means the word frequency in the CELEX database of English. Even though word frequency is not a acoustic feature, it seems that it can aid phonetic classification.



Description: df [15 × 3]

	var1 <chr>	var2 <chr>	cor <dbl>
1	syllstart	syllend	0.9999999
2	syllstart	wordstart	0.9907643
3	syllend	wordstart	0.9907647
4	syllstart	wordend	0.9907647
5	syllend	wordend	0.9907653
6	wordstart	wordend	0.9999998
7	syllstart	tokenstart	0.9907649
8	syllend	tokenstart	0.9907654
9	wordstart	tokenstart	0.9999999
10	wordend	tokenstart	0.9999999

1–10 of 15 rows

Figure 5 couples of features whose correlations are highly correlated in dataset for question 2

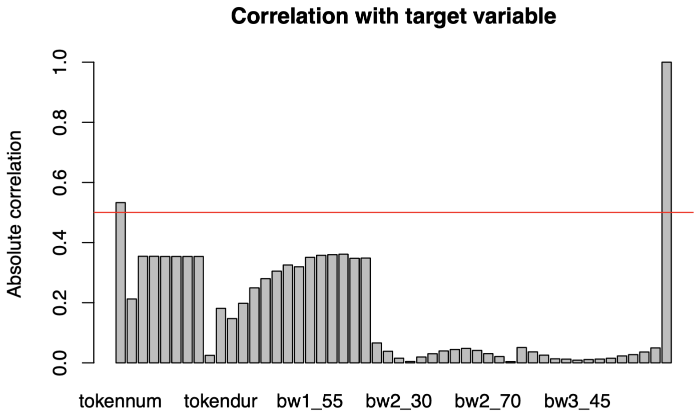


Figure 6 Correlations between features and target in dataset for question 2

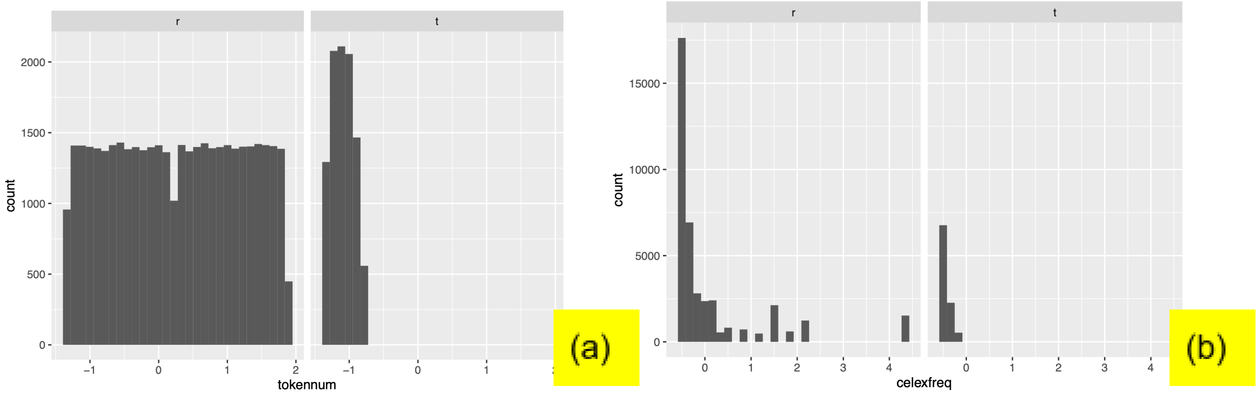


Figure 7 Distribution of the features whose correlations with target are higher than 0.5

(a. tokennum; b. celexfreq)

4.Building and Estimating Models

4.1 Question 1

With pre-processed data, we applied 8 different models for question 1, the 8 models are: Naive Bayes, Linear Discriminant Model (LDA), Quadratic Discriminant Model (QDA), Logistic Regression, K-Nearest Neighbors (KNN), Random Forests, Xgboost and Support Vector Machine (SVM) with 4 different kernels. We chose Naive Bayes as the baseline model and we used confusion matrices to do the evaluations. We did not do the hyperparameter tuning since all the Xs are numerical variables and our computer crashed when we were trying to do hyperparameter tuning.

Table 1 models and hyperparameters chosen

Model	hyperparameter	Accuracy Score
Naive Bayes (Baseline)	/	0.7712
Linear Discriminant Model (LDA)	/	0.7945
Quadratic Discriminant Model (QDA)	/	0.6858
Logistic Regression	/	0.7984
KNN	k=9	**0.8559**
Random Forests	ntree = 1000, maxnodes = 4	0.775

Xgboost	nrounds = 100
	max_depth = 2
	eta = 0.3
	objective = "binary:logistic"
		0.8255

SVM with different Kernels	type = 'C-classification'
	cost = 1
	gamma = 0.5
		Linear: 0.7965
		Polynomial: 0.7124
		Radial: 0.8382
		Sigmoid: 0.6157

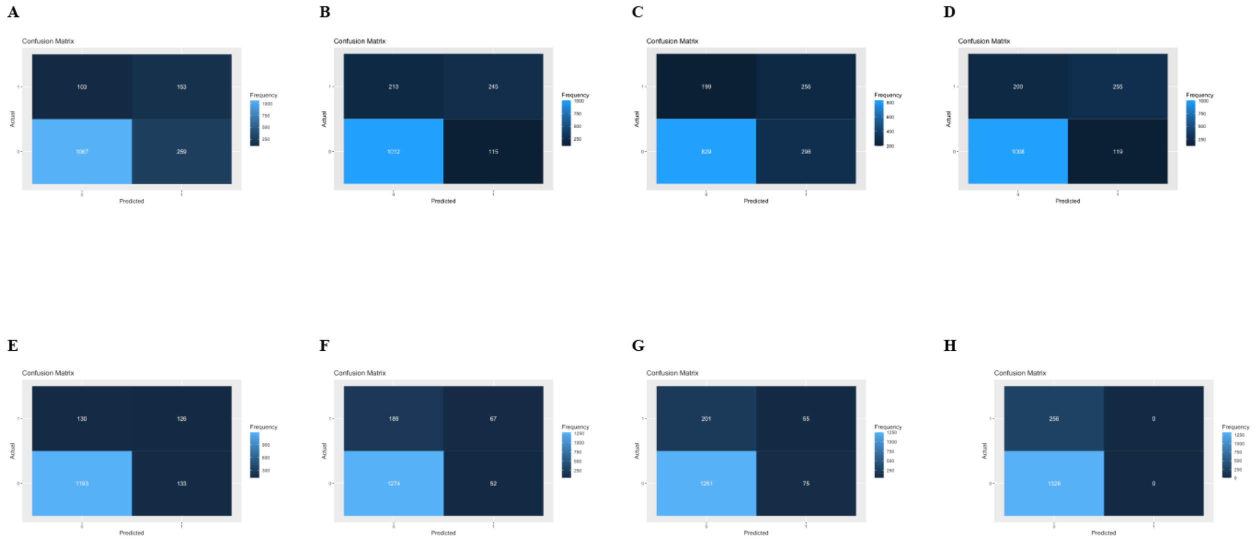


Figure 8 Confusion matrices for 8 different classification models, (A) Naive Bayes, (B) Linear Discriminant Model (LDA), (C) Quadratic Discriminant Model (QDA), (D) Logistic Regression, (E) K-Nearest Neighbors (KNN), (F) Random Forests, (G) Xgboost, (H) SVM RBF.

4.2 Question 2

For question 2, we applied 7 different models: Naive Bayes, Linear Discriminant Model (LDA), Quadratic Discriminant Model (QDA), Logistic Regression, K-Nearest Neighbors (KNN), Random Forests and Xgboost. We also chose Naive Bayes as the baseline model. There exists a data imbalance issue for question 2, the ratio between two labels is around ¼. So, we used F1 Score, ROC curve and Area Under Curve (AUC) Score to do the performance comparison instead of accuracy.

Table 2 models and hyperparameters chosen

Model	hyperparameter	F1 Score	AUC Score
Naive Bayes (Baseline)	/	0.7821729	0.857
Linear Discriminant Model (LDA)	/	1	0.968
Quadratic Discriminant Model (QDA)	/	0.9991	0.999
Logistic Regression	/	0.8843	0.989
KNN	k = 7	0.8377	0.705
Random Forests	mtry = 2	0.3741972	0.615

	importance = TRUE
	ntree = 1000
	maxnodes = 4		

Xgboost	nrounds = 100
	max_depth = 2
	eta = 0.3
	objective = "binary:logistic"
		0.9911474	**1**

** F1 score is a measure of a classification model's accuracy that combines precision and recall. It is the harmonic mean of precision and recall, with a value ranging from 0 to 1, where 1 indicates perfect precision and recall, while 0 indicates the worst possible performance.*

** ROC Curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds, ranging from 0 to 1.*

**The area under the ROC curve (AUC) is a measure of the performance of the classifier, with a value ranging from 0 to 1. AUC of 1 indicates perfect classification, while an AUC of 0.5 indicates the classifier is no better than random guessing.*

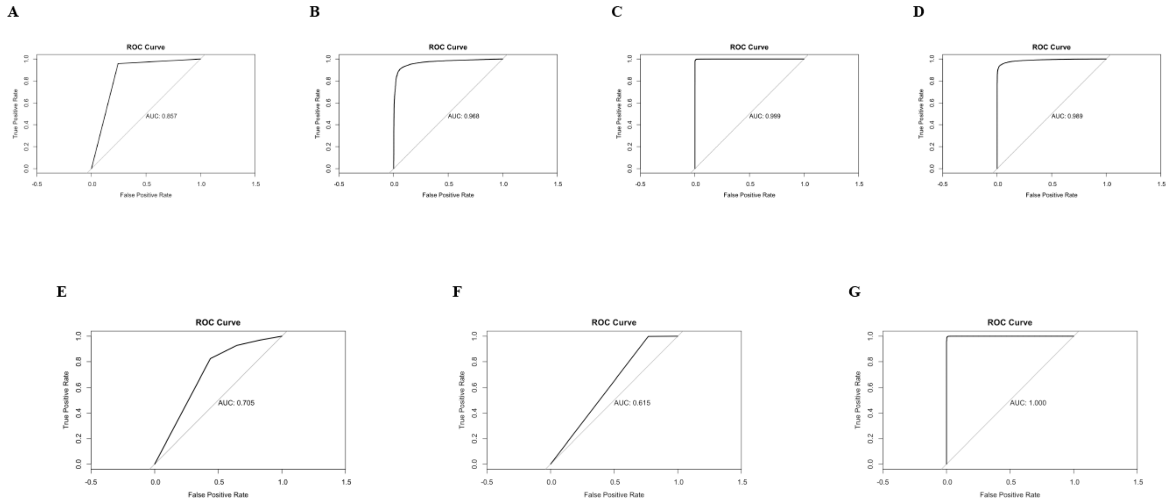


Figure 9 ROC Curve Plots for 7 different classification models, (A) Naive Bayes, (B) Linear Discriminant Model (LDA), (C) Quadratic Discriminant Model (QDA), (D) Logistic Regression, (E) K-Nearest Neighbors (KNN), (F) Random Forests, (G) Xgboost.

5.Conclusion and Discussion

5.1 Conclusion

For question 1 which is to classify the presence or absence of /r/, the best performance model is KNN with the accuracy score equals to 0.8559. Also, the best performance model for question 2, which is to distinguish between /r/ and /t/ is Xgboost with the AUC score equals to 1. The Quadratic Discriminant Model (QDA) also performs well for the task of question 2.

5.2 Discussion

Our question 1, which is to code the presence or absence of /r/, is also the research question in Villarreal et al. (2020). Villarreal et al. (2020) argued that random forest is an effective model to conduct this task, and it has advantages over linear regression or logistic regression. However, our comparison of models suggest that KNN or SVM is outperforming random forest. Therefore, we have made a contribution to the understanding of autocoding models for the /r/ variable.

In order to estimate the performance for the imbalanced dataset of question 2, we used F1 score, AUC score, and ROC curve. The AUC score for Xgboost is 1, which is perfect but a little bit suspicious. However, since the F1 score for the Xgboost model is also very good (0.99), we claim that Xgboost is the best model. From the acoustic point of view, distinguishing between /t/ and /r/ is easier than classifying the presence or absence of /r/. /t/ and /r/ have different places and manner of articulation, and they are also different in terms of voicing. Therefore, it is reasonable to have perfect classification. Classifying the presence or absence of /r/ is more difficult since /r/ has been argued to be a continuous variable. For instance, there could be 5% of /r/, 80% of /r/, etc.

Reference:

[1] Villarreal, D. & Clark, L. & Hay, J. & Watson, K., (2020) “From categories to gradience: Auto-coding sociophonetic variation with random forests”, *Laboratory Phonology* 11(1): 6. doi: https://doi.org/10.5334/labphon.216
[2] https://github.com/nzilbb/How-to-Train-Your-Classifier