Gender Classification on Race: A Speech Perspective Project Report

Harry Xiong, Hank Zhang
xiong1@uchicago.edu, hqzhang@uchicago.edu

Abstract

The current research stems from previous work on speech gender classification and aims to investigate classification bias for African and Asian speakers. 7 models including k-NN, Naive Bayes, SVM, linear perceptron, logistic regression, ANN and CNN are deployed to compare accuracies, False Positive Rates (FPR) and False Negative Rates (FNR) on a data set containing 837, 525 and 189 speaking samples from White, Asian and African speakers. Feature extraction is performed to transform MP3 files to Mel-Frequency Cepstral Coefficients (MFCCs). The results suggest that there is no observable difference in the error rate on the Asian test set compared to the baseline White test set, while the African test set achieved lower classification accuracy compared to the other two.

Introduction

Gender classification tasks in speech have been extensively studied. Various research endeavors have laid emphasis on improving accuracies. The current research, on the other hand, proposes to juxtapose gender classification results of African and Asian speakers and investigate whether any difference exists. The current study is also inspired by a growing body of literature on racial and gender stereotypes in social science fields. Bertrand and Mullainathan [1] in 2005 found that African American-name resumes received less callbacks than white-name resumes. Hall et al. [2] extended the finding to two-dimensional by exploring gender-race interplay in the recruiting process. The current resesearch aims to present difference in gender classification for African and Asian speakers.

Previous Work

Past literatures by Hu et al.[3], Djemili et al.[4], and Lee and Kwak [5] exhibited excellent accuracies above 90% on gender classification using various models such as GMM, multilayer perceptron, SVM, and decision tree. But the data sets used typically do not contain a balanced distribution across races. The current project therefore aims to fill the gap and study the variation in gender classification performance conditioned on races.

Data

A Kaggle data set [6] that contains 2140 same English speech samples of speakers from 177 countries with 214 different native languages is used. Since the races are not specified for the speakers, we have used the countries of origin as a proxy for races. Specifically, speaking samples of speakers from U.S, Canada and Europe are included in the White data set, whereas samples of speakers from Asia and Africa are included in the Asian and African test sets (525 samples and 189 samples) respectively. The White data set is further split into a training set (711 samples) and a test set (126 samples). MP3 files are transformed to Mel-Frequency Cepstral Coefficients (MFCCs) using the librosa library with a sampling rate of 16000 fs. 40 MFCCs are extracted for each sample and data is padded for the CNN model.

Experiments

The experiments will use the MFCC features as the independent variables and the speaker's gender (0/1) as the dependent variable. In order to observe patterns in gender classification performance across races, we will train a number of models so that model-dependent patterns are eliminated.

Determining the Baseline Errors

To arrive at the final models, the model tuning process is carried out on white-speaker data. Eventually, we aim to compare test set errors on 3 test sets, one for each race. Our hypothesis is that results for African Americans will be biased towards misclassifying more females to males relative to the White control group and for Asians it will be the opposite.

To obtain the test set error for White speakers, a test set is held out under a 85-15 train-test split. And the training set will be further partitioned (for 10-fold cross-validation in most models and for simple cross-validation when tuning the CNN model) in order to select the best hyperparameters for each model. Finally, the test set error for White speakers will be used as the baseline error rate for the trained model. The data for African and Asian speakers will be used solely as test sets.

Model Selection and Tuning

We train a total of 7 models in order to ensure any difference in error rate, if any, would be indeed persistent and not coincidental. In the extreme case, performance gaps might even be model-independent, but we do not necessarily anticipate such clean results.

The models that we experiment on will be k-NN, Naive Bayes, SVM, linear perceptron, logistic regression, artificial neural networks (ANN), and convolutional neural networks (CNN). For each model, we tune the hyperparameters, if any, until a validation set (consisting of White speakers) accuracy of near 90% is reached.

Table 1 presents the model details for the first 6 models, and details for the CNN model can be found in figure 1.

Models	Model Details
k-NN	$k = 8$
Naive Bayes	Gaussian Naive Bayes
SVM	degree-3 polynomial kernel L2 regularization: $\alpha = .1$
Perceptron	linear perceptron L2 regularization: $\alpha = .00005$
Logistic regression	L2 regularization: $\alpha = 10$
ANN	1 hidden layer of 100 nodes tanh activation

Table 1: model details

CNN-based Model

For the CNN model, we tuned the model based on validation set performance while fixing the model layers fixed. We were able to find the optimal number of channels to include in each convolutional layer to be 16 and 32 respectively. The convolutional layers are followed by a fully connect layer and finally a softmax layer. The model achieved over 97% accuracy on the validation set and as we will see, around 93% accuracy on the test set.

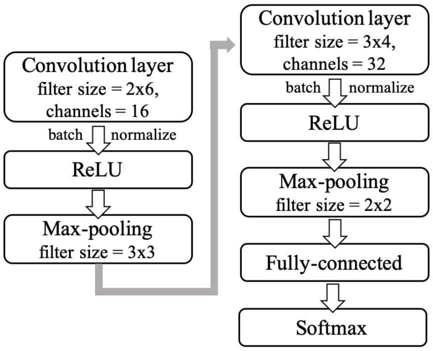

Figure 1: CNN model diagram

Results

After using the trained models on the 3 test sets, we are able to observe the difference in error rates. It would not be surprising if the test set error rates are higher on both the Asian and the African test sets because after all, the models were tuned on the White training and validation sets.

Interestingly, what we observe is that in general, there is no observable difference in the error rate on the Asian test set compared to the baseline White test set. For some models, the performance on the Asian test set is even better. On the other hand, all models predict less accurately on the African test set, compared to on the Asian counterpart. Further, except for the SVM model, the test set accuracy is lower for African speakers compared to the baseline White speakers.

For detailed test set accuracies, please see table 2. Further breakdown of the precisions, recalls and errors for the first 6 models in false positive rates (FPR) and false negative rates (FNR) are provided in the Appendix.

Models	Baseline	African	Asian
k-NN	.825	.730	.787
Naive Bayes	.889	.873	.901
SVM	.889	.910	.939
Perceptron	.825	.788	.842
Logistic regression	.929	.910	.928
ANN	.913	.910	.935
CNN	.937	.878	.930

Table 2: test set accuracies

Our initial hypothesis that the FPR for African speakers would be higher than the Asian speakers and that the FNR would be the opposite is not confirmed by the experiment results. Specifically, compared to Asian speakers, both FPR and FNR are higher on the African test set. Nevertheless, the consistently higher error rates on the African test set across all models is a strong indication that gender classification may be conditioned on race.

Discussion

There is room for improvement in terms of the data set used. Our data set does not have information on race, and countries of origin can be an inaccurate proxy.

Secondly, we are not able to control for accents. Ideally, the speakers in the data set should all be native English speakers of a single type of accent. Otherwise, phonetic differences could in fact be the reason causing the lower accuracy on the African test set.

Arguably, the gender balance between male and female speakers can be improved. However, since the FPR and FNR are mostly both higher on the African test set, gender balance is unlikely to have caused biased error rates in our experiments.

Overall, for the consistently lower prediction accuracy on the African test set, our study suggests that training gender classification models on predominantly White English speakers could cause undesirable performance issue for African speakers. This is an indication that gender classification is indeed conditioned on race. Consequently, social implications could be further explored to potentially address the gendered stereotypes in race.

References

[1] M. Bertrand and S. Mullainathan, “Are Emily and Greg more employable than Lakisha and Jamal? A field experiement on labor market discrimination,” *American Economic Review*, 94.4 (2004): 991-1013.

[2] E. Hall, A. Galinsky, and K Phillips, ”Gender profiling: A gendered race perspective on person-position fit,” *Personality and Social Psychology Bulletin*, 41.6 (2015): 853-868.

[3] Y. Hu, D. Wu, and A. Nucci, “Pitch-based gender identification with two-stage classification,” *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.

[4] R. Djemili, H. Bourouba, and M. C. A. Korba, “A speech signal based gender identification system using four classifiers,” in *Proceedings of the 2012 International Conference on Multimedia Computing and Systems*, pp. 184–187, Tangiers, Morocco, May 2012.

[5] M.-W. Lee and K.-C. Kwak, “Performance comparison of gender and age group recognition for human-robot interaction,” IJACSA International Journal of Advanced Computer Science and Applications, vol. 3, no. 12, 2012.

[6] Kaggle Speech Accent Archive. Parallel English speech samples from 177 countries. *https://www.kaggle.com/rtatman/speech-accent-archive?select=speakers_all.csv*.

Appendix

Models	**Baseline**	**African**	**Asian**
k-NN	.766	.794	.741
Naive Bayes	.855	.897	.892
SVM	.867	.917	.933
Perceptron	.936	.975	.972
Logistic regression	.909	.931	.925
ANN	.894	.946	.940

Table 3: test set precisions

Models	**Baseline**	**African**	**Asian**
k-NN	.937	.761	.858
Naive Bayes	.937	.897	.917
SVM	.921	.940	.941
Perceptron	.698	.675	.692
Logistic regression	.953	.923	.925
ANN	.937	.906	.925

Table 4: test set recalls

Models	**Baseline**	**African**	**Asian**
k-NN	.286	.319	.279
Naive Bayes	.159	.167	.103
SVM	.143	.139	.063
Perceptron	.048	.028	.018
Logistic regression	.095	.111	.070
ANN	.111	.083	.055

Table 5: test set false positive rates

Models	**Baseline**	**African**	**Asian**
k-NN	.063	.239	.119
Naive Bayes	.063	.102	.083
SVM	.079	.060	.059
Perceptron	.302	.325	.308
Logistic regression	.048	.077	.075
ANN	.063	.094	.075

Table 6: test set false negative rates