ENCS5344 Spoken Language Processing Course Project

Dana Assad
Department of Electrical and Computer Engineering, Birzeit University
1211452@student.birzeit.edu

Abstract: In this paper, we will describe a system for recognizing the ethnic origin of British speakers as Asian or White based on the analysis of MFCCs of speech samples. In our case, the objective of the system is to classify speech segments into one of two ethnicity classes. With the help of a dataset containing .wav files of British speakers. The proposed system was quantitatively assessed with the use of training and testing sets which were defined in advanced and the accuracy was used as the measuring parameter. These findings show how the suggested framework can be applied to ethnicity recognition, and indicate that the highest outcomes were obtained by tuning the hyper-parameters for the KNN and SVM algorithms. These papers are useful to provide effective recommendations on the use of acoustic features and machine learning for demographic classification problems.

Index Terms: MFCCs, Speech Recognition, Speech processing, Ethnicity Classification, Machine Learning Algorithm (SVM, KNN, and GMM).

I. Introduction

The possibility to distinguish the speaker by their gender, age, ethnicity, and other characteristics is extremely valuable for sociolinguistics, speech recognition and speech technology, human-computer interaction, and so on. In this project the emphasis is put to differentiate between two particular ethnicity therefore between Asian and White ethnicities when it comes to their speech. The selection of these groups allows to investigate the anonymity of disturbances created by sociocultural and linguistic factors, disclose the phonetic peculiarities typical for each group.

The ethnicity identification problem poses several problems inherent in the field of speech processing including variations in speaking pattern, accent, and noise levels. Further, the applicability of such systems is in real scenarios where the flexibility of self-paced, natural, and clear speech is inevitable, and the system should be very accurate.

The main goal of this project is to develop and test an approach for obtaining relevant information from samples of speech using Mel Frequency Cepstral Coefficients (MFCCs) and its related acoustic features. The data is then fed into machine learning models K-NN and SVM in order to categorize these features into one or the other target ethnicities. This work enriches the theoretical knowledge on speech processing technologies and helps in implementing related tasks including demographic assessment, designing speech interfaces suitable for specific users, and improving individualization in Voice UI.

In order to achieve these goals, the project is designed with extended experiments with a dataset of the recorded speech samples of British speakers of both ethnicities. Accuracy, precision, recall and F1 score are used as the metrics to determine the effectiveness of the ethnicity classification, with performance data indicating the efficiency of the suggested system.

II. Background and related work

The identification of ethnicity based on speech has attracted lot of interest in speech processing and has many uses in sociolinguistic studies, individualization of speech-based systems, and demographic profiling. Various aspects of speech such as MFCCs have been adopted commonly for speech recognition because of their capabilities to describe the spectral characteristics of the sound. MFCCs are a method that was introduced by Davis and Mermelstein in 1980 and they are based on the human peripheral auditory system; making them effective for ethnicity classification due to their ability to capture small variations that exist in phonemes [1].

Several speech classification methods are used where SVMs and KNNs are common machine learning techniques. SVMs are particularly popular for their ability to deal with non-linear classification and as a result can pin down more intricate patterns in speech data [2].

As with the KNN classifier, there is ease and efficiency when defining clear classification decision boundaries can be efficiently met. Both models offer accurate architectures for segmenting and modeling sounds' features and detecting demographical characteristics in speech.

However, several difficulties exist regarding ethnicity recognition by speech. Uncontrolled pronunciation of speakers, noisy environment, and scarcity of extensive datasets are some of the challenges faced [3]. These difficulties mean that complex methods of feature extraction and such flexible classification techniques as the modified versions of the machine learning algorithms should be applied to address the issue of accommodation to various variations in the spectrum of a speech signal.

Recent research have shown that ethnic and accent identification is another area in which machine learning can be effectively applied. To further prove that demographic characteristics are identifiable from Acoustic features, Abu El-Haija et al. (2020) go on to show the viability of MFCCs for distinguishing regional Arabic accents. Furthermore, Najafian et al. (2020) pointed out that deep learning models offer higher classification performance for tasks with varying and intricate present the benefits of deep learning models in enhancing classification efficiency when working with complicated speech materials. These studies highlight the need for addressing the challenges in linking higher forms of analysis with the basic speech demographic classification [4].

In this project, MFCCs are utilized alongside SVM and KNN models to classify British speakers into two ethnic categories: Asian and White. Thus, this research examines ethnicity recognition from the aspect of acoustic features and applies machine learning to fill the gap in the research issue and make valuable contributions to the development of speech processing and demographic analysis.

III. Methodology

To achieve this for this project the following steps are followed in order to create a machine learning model for ethnicity classification between Asian and White Britons. It consists dataset preparation, feature extraction, model training, hyper-parameter tuning, and performance evaluation. Each stage is detailed below:

A. Dataset Preparation

The dataset used for this study comprises speech recordings of British speakers belonging to two ethnic groups: Asian and White. This was done to make the tests conducted on the model conducted more fairly and have a training set of dataset and a testing set for the same. There are several initial steps in cleaning of the audio that include: background noise reduction, silence, and normalization of the raw unprocessed files. Files contain audio in the WAV format.

⇨ Python script was designed with features necessary for preprocessing of audio samples data. ***Librosa*** performed the load, resample, and remove noise, While ***Pydub*** split and trim silence of the segments. ***Scipy*** wrote the processed files, and ***noisereduce*** enhanced the tones. In conjunction, they preprocessed the dataset to allow it to go through various applications of machine learning.

B. Feature Extraction

The main feature extraction method utilized is Mel Frequency Cepstral Coefficients (MFCCs), representing the short-term power spectrum of sound using a nonlinear Mel frequency scale.

$$\boldsymbol{MFCCs} = \sum_{\boldsymbol{n=0}}^{\boldsymbol{N}} \boldsymbol{log}\left(\boldsymbol{S(n)}\right) \boldsymbol{cos}\left[\frac{\boldsymbol{K}\left(\boldsymbol{n}-\frac{\boldsymbol{1}}{\boldsymbol{2}}\right)\boldsymbol{\pi}}{\boldsymbol{N}}\right], \quad \boldsymbol{k = 1, .., k}$$

➔ Where S(n) is the power spectral density of the signal, N is the number of filters in the mel-filter bank, and is the number of cepstral coefficients. This project utilizes 20 MFCCs per audio frame, along with their first and second derivatives (delta and delta-delta coefficients), including energy and ZCR to capture dynamic changes in the features.

⇨ It was performed using the librosa library in Python for fast and normalized result extraction for all files.

C. Model Training

For the classification task, two machine learning models were utilized: K-Nearest Neighbors (KNN) which classifies data based on majority of the nearby points, where K is tuned as the hyper parameters Values for K were chosen by hand based on the intuition that it should be small given the structure of our dataset. SVM which classifies data based on hyperplane separating them in higher dimensions with the help of tuning regularization parameter(C) and kernel type respectively was tuned using Both of them are experimented to know that how much their classification performance is effective in recognizing ethnicity.

D. Model Testing

In the analysis of the outcome of the models, various evaluation metrics were used, namely accuracy, precision, recall and F1 score. Having created the confusion matrices, the classification results could have

been examined in detail. These metrics offered measures which included the percentage accuracy in identifying the two groups of people belonging to the two ethnic groups and for the different models.

IV. Experiments and Results

a) Experimental Setup:

All the experiments are implemented using Python, and machine learning libraries including Scikit-learn and Librosa for feature extraction. The results of the experiments, as well as effects of changing the hyper-parameters, were plotted using the same matplotlib library for better analysis.

Our dataset consists of labeled audio samples from two regions: Asian and white.
Each class in the dataset contains audio files that have been converted into MFCC features. The training dataset includes samples from each accent class, while the testing dataset is used to evaluate the model's performance.

b) Model Configuration

For this project, three different **machine learning models,** including Support Vector Machine (SVM), K-Nearest Neighbors (KNN) as well as Gaussian Mixture Model (GMM) are used to distinguish the origin of speech of the British people. The SVM was optimized through GridSearchCV, adjusting hyperparameters like kernel type (Linear and RBF) and the regularization parameter (C: 0.1, 1, 10, 100). For KNN, the Hard Coded value of numbers of neighbors were taken as 3, 5, 7 and 9 based on Euclidean Distance. The GMM was trained with two components and tested with different types of covariances (full, tied, diag) to resolve the problem of classification.

c) Evaluation Metrics

Model performance was evaluated using the following metrics: Measures of performance include: –Accuracy –Precision –Recall –F1 score. These metrics were selected in order to have a comprehensive evaluation of the model's performance.

d) Results

Figure-1 presents the accuracy performances of the classification models; SVM, KNN, and GMM with different settings and parameters such as setting of C (Regularization parameter) of SVM, setting of k (number of neighboring points considered) of KNN and covariance type of GMM. The performance of each model for classification of the speech data is compared using accuracy, precision, recall, and F1 measures for each of the configured data sets.

Summary of Results:

	Model	Accuracy	Precision	Recall	F1 Score
0	KNN (k=3)	0.650	0.656250	0.650	0.646465
1	KNN (k=7)	0.625	0.642450	0.625	0.613153
2	KNN (k=9)	0.675	0.719436	0.675	0.657670
3	SVM (C=0.1, kernel=linear)	0.675	0.699430	0.675	0.664732
4	SVM (C=1, kernel=linear)	0.700	0.738095	0.700	0.687500
5	SVM (C=10, kernel=linear)	0.650	0.678571	0.650	0.635417
6	SVM (C=100, kernel=linear)	0.625	0.642450	0.625	0.613153
7	GMM (covariance_type=full)	0.500	0.250000	0.500	0.333333

Figure 1: Performance models result

→ Hyperparameter values also has an effect on the accuracy of the models which is plotted below in **log scale.**
The graph represents the occurrence of classification accuracy for both types of models, according to the values assigned to hyper-parameters C for SVM and K for KNN. The SVM with linear kernel learns exemplar more accurately than RBF while the accuracy increases gradually with increasing of C and starts to level-off after moderate level of C On the other hand, the SVM using the RBF kernel has the lowest accuracy across the board for all the values of C, The graph of the KNN model illustrates that when the value of K varies, its accuracy increases and decrease at a certain extent. This gives credit to the fact that accuracy varies with the hyper-parameters of each model.

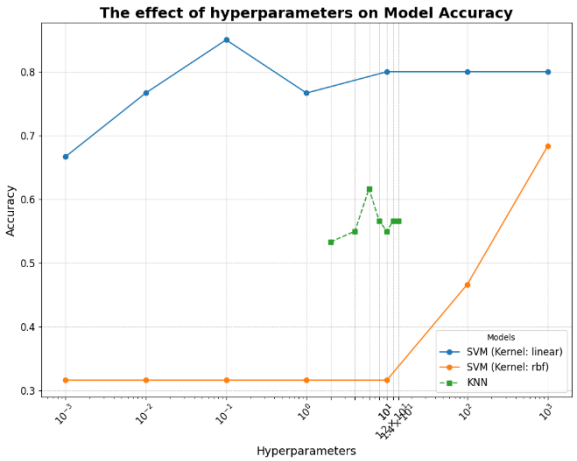

Figure 2: The hyper-parameter Effect

⇨ Confusion matrices
The results reflected in the confusion matrices of the assessed models, which included SVM, KNN, and GMM, were also indicative of differences in classification accuracy. A linear kernel SVM with (C = 1) achieved the best overall discriminative capability and minimized the confusion between "Asian" and "White" faces, with Accuracy=**70%.**

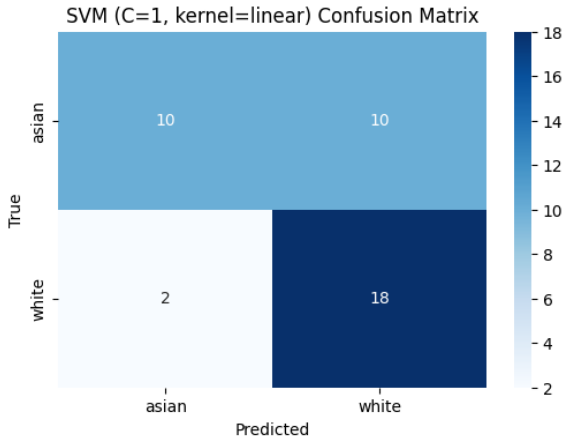

Figure 3: SVM Confusion Matrix

Likewise, there was an enhanced classification with KNN that predicted with (k = 9) boosted the results as compared to other K values predicted. With Accuracy = **68%.**

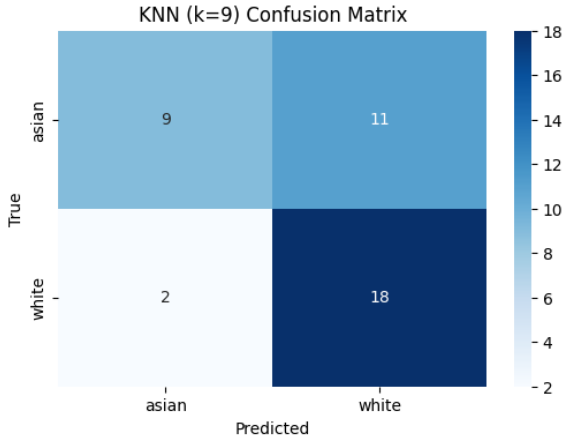

Figure 4: 9-NN Confusion Matrix

However, the GMM model seemed to have some issues with the separation of two classes, which was closely related to the choice of model and its hyper-parameters. Its accuracy = 50%

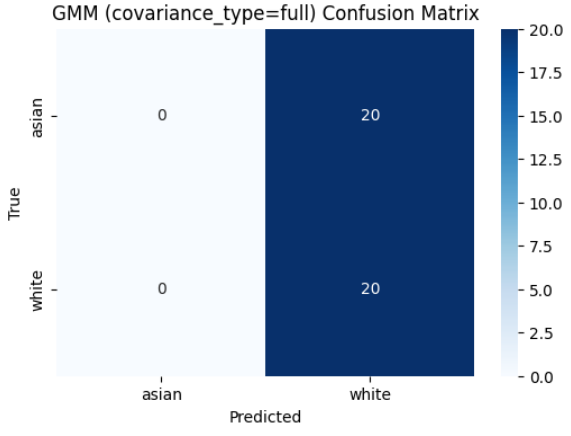

Figure 5: Full GMM Confusion Matrix

e) Discussion

The experiments show that difference in configuration and preprocessing yields distinct model performance. SVM with specified parameters as \(C = 1\) and the linear type of kernel was the best performing with the highest accuracy of 70% but with sufficient model complexity. **KNN with \(k = 9\)** is also good enough that demonstrated the significance of neighbors' selection. On the other hand, the **GMM model** was not very successful, which suggested its inaptness for the data without modifications. This particularly was achieved by normalizing features which led to low accuracy below 50% due to their representation of key relations that might have been derailed by normalization. This underlines the importance of critiquing preprocessing effects on raw data information for pertinent data to be captured for useful speech classification.

V. Conclusion and Future Works

In this project, we have been able to apply and test several machine learning models, such as SVM, KNN, and GMM, for Speech origin classification of speech data into "Asian" and "White" subclasses. The accuracy values established showed that the SVM with linear kernel and (C = 1) provided the optimal accuracy to address this classification task. KNN with (k = 9) also devised reasonable accuracy which can achieve reasonable accuracy but GMM failed to map the classes properly because of the problem of feature extraction in this paper, highlighting the understanding of model selection and hyper-parameters tuning.

The future work may target toward collection of larger samples from different population for testing the proposed approach as well as implementation and comparison of deeper models like ensemble techniques or the use of deep learning paradigms. Moreover, a deeper exploration of preprocessing techniques, as more advanced feature scaling or feature ranking techniques, could enhance the achieved results as well. Domain knowledge and acoustic analysis may also strengthen the accuracy of classification and accelerate the development of more efficient systems of speech recognition.

VI. Partners participation tasks

The responsibilities of the project were handled on my own as I worked autonomously on every element. I took full responsibility for composing the Abstract, Background, Introduction, Methodology, and Conclusion. Furthermore, I carried out the Experiments and gathered the Results completely by myself.

.

VII. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366,1980.

[2] W. M. Campbell, D. E. Sturim and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," in *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, May 2006.

[3] Schultz, T., & Waibel, A. (2001). "Language-independent and language-adaptive acoustic modeling for speech recognition". *Speech Communication*, 35(1-2), 31-51. (Language-independent and language-adaptive acoustic modeling for speech recognition - ScienceDirect)

[4] A. Abu El-Haija, N. Madi, I. Mustafa, N. Abed, M. Afana, and A. Abuzaina, "Palestinian Arabic regional accent recognition," Birzeit University, 2020

[5] M. Najafian, M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Accent recognition with deep neural networks: Analysing effect on different levels of granularity," Journal of Signal and Information Processing, vol. 11, pp. 45 56, 2020,

VIII. Appendix

Project Code in this Link:
https://colab.research.google.com/drive/17FQrVr85jZAM2wFqNqniBzUWXbEI5ibM?usp=sharing

Or in "Code_SPL.txt"