Develop and Evaluate a Recognition System of Ethnic Groups From Speech.

ENCS5344 - SPOKEN LANGUAGE PROCESSING, JANUARY 2025

Leena Affouri - 1200335, Mariam Hamad - 1200837, Salwa Fayad - 1200430
Department of Electrical and Computer Engineering
Birzeit University
Ramallah, Palestine
Supervised by: Dr. Abualsoud Hanani

***Abstract*—There are many accents in the UK, especially in Birmingham city based on the people's ethnicity, but there are two main ethnic groups Asian and White British residents. This project develops a speech system that classifies speakers based on their ethnicity either Asian or White. We implemented three techniques and adopted several sound features to achieve this goal. Features used is energy, zero-crossing rate, pitch frequency and 12 Mel-Frequency Cepstrum Coefficients (MFCCs) with their deltas and delta-delta and the three machine learning models is K-Nearest Neighbors (KNN), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM). These models were trained and tested on dataset containing speech for both males and females for each accent speakers. Then, the system performance measures accuracy, precision, recall, F1-score, and the confusion matrix for each model.**

***Index Terms*—MFCCs, UK accents, accent recognition, speech processing, KNN, GMM, SVM.**

I. Introduction

In recent years, there has been increased interest in the process of recognizing the human race through speech by analyzing the phonetic features extracted from spoken language. This project aims to develop a system for recognizing race, specifically designed to distinguish between two major ethnic groups in the city of Birmingham, UK, which contains many different ethnic groups, with the majority of the population being ”white” British and ”Asian” English speakers. The project aims to create a system that can discriminate and identify these groups based on speech features, utilizing the ”Voices Across Birmingham” corpus. It contains phone recordings between city residents as data for the voices of both ethnic groups, making it ideal for training and testing a race classification system. The goal is to enable automatic and effective ethnicity classification from voice recordings. Using speech characteristics and machine learning techniques, the system will categorize speakers into one of two ethnic groups. The models will be trained and tested to determine how accurate they are in differentiating between the two ethnic groups.

II. Background/Related Work

Speech recognition systems it's a process to understand and interpret human speech then convert it into text, it become very important over the past years because it depend on sophisticated algorithms and machine learning to understand the human speech in real time even with the difference in accents, pitch, speed, and slang. It's used in different fields such as Virtual Assistants like Siri, Alexa, and Google Assistant and Accessibility Tools these are special apps for help disable people and other application like health and customer service [1].

Over the years, these was a lot of effort made to develop methods for identify speaker attributes, including ethnicity, accents, and dialects, from speech. much research has been made on this exploring different models based on acoustic features, which later also included deep neural networks. Starting from 2000, research done by Reynolds and Quatieri, who introduced a Gaussian Mixture Model-based model (GMM) and universal background model (UBM) for speaker detection and tracking in multi-speaker audio, the paper evaluates how various components of the detection and tracking algorithms influence the overall system performance [2].
In 2005, Yanli Cheng and his team do study about the accent detection and speech recognition for Shanghai-Accented Mandarin in this study they combine Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) adaptation, optimizing their integration for improved recognition systems in multilingual and multi-accent contexts [3].

In 2006, Konstantin Markov and Satoshi Nakamura they developed a hybrid Hidden Markov Model (HMM) that use a Bayesian Network (BN) model to represent the state probability distribution, instead of the use of Gaussian mixture models which is commonly used [4]. In the same year William Campbell and Douglas E. Sturim they explore the idea of using a super GMM in a support vector machine (SVM) classifier. They combine the recent results in SVM methods with the GMM supervector concept [5].

In 2011, A. N. Mishra and his team in their study they develop a recognition system for connected Hindi digits using different features. their main work is to show that the recognition performance of connected Hindi digits using Hidden Markov model Toolkit (HTK) with Mel-frequency cepstral coefficient (MFCC) and other feature extraction techniques like Perceptual linear prediction (PLP), Revised PLP, Bark frequency cepstral coefficients (BFCC) and Mel frequency PLP (MF-PLP). The experiments were conducted on clean as well as on noisy data for connected Hindi digit recognition [6]. In the same yeas, study by Biadsy, Fadi their study was on different Arabic and English accents they combine different features and approaches that allows the study to achieve state-of-the-art performance in dialect and accent recognition while improving ASR systems [7].

In 2013, Santosh Kashinath Gaikwad, Dr.Bharti W Gawali and Karbhari Kale they present an experimental approach of acoustic speech feature for Marathi and Arabic accents for English speaking. using formant frequency, energy, and pitch characteristics features [8].

In 2016, Pham Ngoc Hung and his team they combine different features to improve the recognition of Vietnamese dialects. By combining MFCC coefficients with fundamental frequency (F0) information, the recognition performance improves from 58.6% to 70.8% [9]. at the same year another group consist of Yishan Jiao, Ming Tu, Visar Berisha and Julie M Liss they do accent identification by combining Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) trained on long- and short-term features [10].

III. METHODOLOGY

After comprehending the idea of the project and analyzing the dataset, which includes the differentiation of two ethnic groups – 'Asian' and 'White'- based on their English speech, we decided to follow a specific structure. All the steps involved in this methodology are very important in meeting the goals of the project as well as the right classification and evaluation. The overall methodology of the research is summarized in a block diagram in Figure 1.

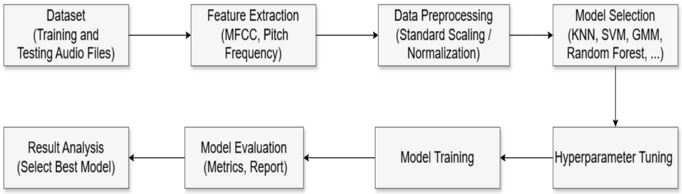
Dataset (Training and Testing Audio Files)
Feature Extraction (MFCC, Pitch Frequency)
Data Preprocessing (Standard Scaling / Normalization)
Model Selection (KNN, SVM, GMM, Random Forest, ...)
Hyperparameter Tuning
Model Training
Model Evaluation (Metrics, Report)
Result Analysis (Select Best Model)

Fig. 1. Block diagram for the Methodology.

The primary objective of this project is to classify speech samples into two categories: The authors compare the 'Asian' and 'White' English speakers. To achieve this we followed a systematic approach whereby we extracted audio features, preprocessed the data, used machine learning algorithms and finally evaluated our model. The next subtopics explain the process followed and the models used in classification in details.

A. Data Preprocessing

In the first, our dataset is in .WAV format. It's organized into separate training and testing data folders, with each folder subdivided by ethnic group—into Asian and White speaker files. Also, within the training folder, each speaker is divided by gender into subfolders for male and female speakers separately. To ensure that the training process is more accurate and evaluation models improve overall performance and precision of the classification, the dataset of each ethnic group is processed separately.

Then in the data preprocessing part, we remove the silences from all audio files to enhance and increase the accuracy and performance for the models by pointing out the segments with speech and discarding the silent parts. So these silent removal processes were made using the librosa library and the librosa.effects.split() function that is used to detect non-silent intervals based on a threshold top_db, and the silent parts are discarded. Once we have identified non-silent intervals, all the non-silent audio segments are concatenated to make the final non-silent audio signal.

Then after removing silents, we make feature extraction. In this case, we extract various audio features that are helpful in discriminating between the different ethnic groups. The features include [12]:

1) **Mel-Frequency Cepstral Coefficients (MFCC)**

   MFCCs are widely used in speech recognition to represent the power spectrum of a signal in the models. They record the frequency properties of the sounds to assist in discerning between distinct speech patterns.

   The MFCC is usually obtained from the Fourier transform of the audio signal using the librosa.feature.mfcc() library and involves windowing, Fourier transform, Mel scale transformation, and DCT (Discrete Cosine Transform) steps. The windowing is used to divide the audio signals into overlapping frames by using a window function such as a Hamming window, then apply the Fourier transform for each window to obtain the frequency spectrum, and then these frequency scales are transformed into Mel scale. This corresponds to human hearing sensitivity by using Mel Scale Transformation; in the last step, apply DCT to the mel spectrum to obtain the MFCCs.

   The MFCC equation is:

$$\text{MFCC}(k) = \sum_{n=1}^{N} \log\left(\|X(n)\|\right) \cdot \cos\left(\frac{\pi k(n - \frac{1}{2})}{N}\right) \tag{1}$$

   So the X(n) is the log Mel spectrum, k is the coefficient index, and N is the number of frequencies.

2) **Delta MFCC (ΔMFCC) and Delta-Delta MFCC (ΔΔMFCC)**

   The delta and delta-delta of MFCCs are the first and second derivatives of the coefficients, respectively. It uses the librosa.feature.delta() function. These are useful for capturing modulations of voice, such as changes in pitch.
   The Delta MFCC:

$$\Delta\text{MFCC}(k,t) = \text{MFCC}(k,t) - \text{MFCC}(k,t-1) \tag{2}$$

   The Delta-Delta MFCC:

$$\Delta\Delta\text{MFCC}(k,t) = \Delta\text{MFCC}(k,t) - \Delta\text{MFCC}(k,t-1) \tag{3}$$

3) **Zero-Crossing Rate**

   It's used to measure how many times the signal crosses zero per unit of time or each sample to get the number of times that the signal changes its sign as the result. Which represent features of voice like pitch. Also, it used librosa.zero_crossings() library to apply it.

The Zero-Crossing Rate Equation:

$$\text{ZCR}(t) = \sum_{n=1}^{N-1} \text{sign}(x(n)) \cdot \text{sign}(x(n+1)) \tag{4}$$

Where x(n) is the signal at sample.

4) **Energy**

Is a measurement of the signal's energy that indicates the volume of speech. It used librosa.feature.rms() library to apply it.

The Energy Equation:

$$\text{Energy}(t) = \sum_{n=1}^{N} x(n)^2 \tag{5}$$

Where x(n) is the audio signal.

5) **Pitch**

The pitch of the signal represents the fundamental frequency of the speech. It will help recognize between distinct accents or speakers.
The Pitch Equation:

$$\text{Pitch}(t) = \text{mean of the highest peak frequency at each time step} \tag{6}$$

B. Machine Learning Algorithms

In order to determine which machine learning model gives the highest accuracy when it comes to discriminating between 'Asian' and 'White' English speakers, a number of machine learning classifiers were used. These are the brief description of the models, their methodologies, and any relevant mathematical details:

- **K-Nearest Neighbors (KNN):**

KNN is easy to interpret and understands and it classifies samples by assigning the most frequent class among the k-nearest neighbors. KNN is advisable for use with small and medium-sized datasets, it works well when there is a linkage between the features [11]. The KNN algorithm is implemented using Euclidean distance metrics to locate the nearest neighbor. The Euclidean distance metrics d(x,y) between two points x and y is calculated using Equation (7).

$$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{7}$$

- **Support Vector Machine (SVM):**

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [11]. This model suitable for the high dimensional data. It attempts to identify that hyperplane which separates the two classes most effectively (Asian and White). Representation of linear support vector classifier is as shown in Equation (8):

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{8}$$

- **Random Forest Classifier:**

Random Forest is a technique of a group of decision tree learning algorithm that builds a multitude of decision trees at a time so as to get better precision and avoid overfitting. The hyperparameters were tuned using GridSearchCV to experiment with different values of $n_estimators, max_depth, min_samples_split, bootstrap$and so on.

- **Gaussian Mixture Model (GMM):**

GMM a statistical model that models data points as coming from a number of Gaussian distribution mixtures. This model was adjusted to employ a full covariance matrix and to fit the feature distribution in order to discover clusters, which would correspond to different ethnic groups.

- **Gaussian Naive Bayes (GNB):**

GNB is a simple probabilistic classifier which follows the Bayes Classifier theorem. It has the capability to exploit the fact that the features used are conditionally independent of each other given the class label. This model is ideal when the features are normally distributed and as simple as this model maybe, it can be very powerful for speech classification.

IV. Experiments and Results

This section demonstrates the dataset used, the experiments performed, how they were accomplished, and the results obtained will be explained. In this experiment we classify speech samples into two ethnic groups: "Asian" and "White."

A. Dataset

The dataset contains audio files that separated into training and testing, and each one also divide into Asian and White Table I shows more details about the dataset, the duration for all files for each type their duration is approximately close, while they have the same sample rate of 8000 Hz, ensuring a consistent resolution across all samples. It appears that it is a relatively medium to small data set in terms of the number of speakers.

TABLE I
Dataset Summary

Dataset	Duration (s)	Sampling Rate (Hz)	Signal Length (Samples)	Frames (Seconds)
Testing - Asian	11256.78	8000	90054240	11256.78
Testing - White	24268.90	8000	194151200	24268.90
Training - Asian - Female	29432.22	8000	235457760	29432.22
Training - Asian - Male	14588.39	8000	116707120	14588.39
Training - White - Female	15917.84	8000	127342720	15917.84
Training - White - Male	12425.83	8000	99406640	12425.83

B. Machine Learning Models

Five machine learning models have been used and compared among them to evaluate the results and choose the best two among them. Table II presents the performance metrics of the five models. Based on the results shown in the table, the KNN with (k=5) model shows the best testing accuracy (82.50%) with good precision, recall, and F1 score, making it the best choice to identify the speakers for the project. The SVM (Tuned) model shows the second best performance with high training accuracy (93.33%) and acceptable testing accuracy (75.00%). These two models that we selected for optimization and evaluation due to their overall reliability and balanced performance across metrics. Also shown the result of Precision, Recall and F1 Score, the precision indicates how accurate the positive prediction because it regards the fraction of accurate positive predictions out of all positive predictions generated by the model. Recall shows the ability of the models to identify all relevant cases. Also, the F1-score is the harmonic mean of Precision and Recall.

TABLE II
SUMMARY OF RESULTS

Model	**Training Accuracy (%)**	**Testing Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
K-Nearest Neighbors (k=5)	80.00	82.5	0.8258	0.825	0.8249
Support Vector Machine (Tuned)	93.33	75.0	0.7500	0.750	0.7500
Random Forest (Tuned)	100.00	52.5	0.7564	0.525	0.3866
Gaussian Mixture Model (Full Covariance)	33.33	50.0	0.7500	0.500	0.3333
Gaussian Naive Bayes	64.44	52.5	0.5251	0.525	0.5247

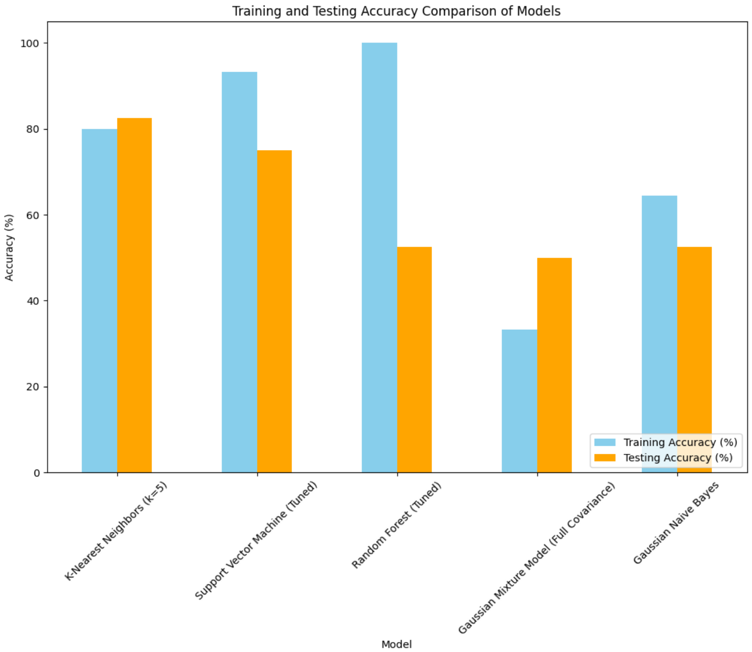

Fig. 2. Training and Testing Accuracy Comparison of Models.

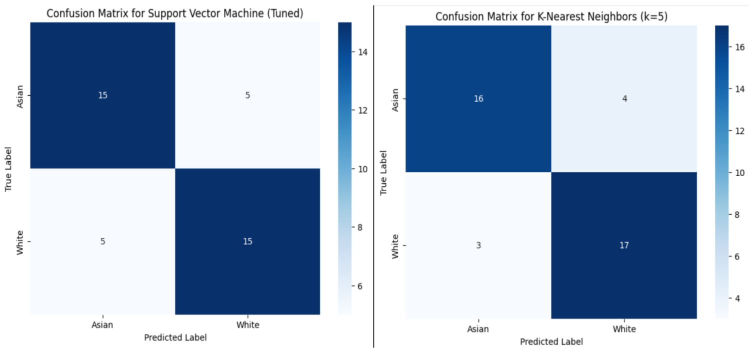

Fig. 3. Confusion Matrix for KNN and SVM.

Figure 3 represents the confusion matrix for KNN and SVM. It's shown on the first row are the True Positive (TP) and the False Positive (FP), while the second row shows the False Negative (FN) and True Negative (TN). For KNN for TP, it predicts correctly 16 times, and for FP predict incorrectly 4 times; in other words, it predicted it as Asian, but it's actually White, and for FN, it predicted it incorrectly 3 times, which means predicted as White, but it's actually Asian, and TN, it predicted correctly 17 times, which means true White predictions, correctly predicted as White, and this is the same for the SVM but with a little difference: TP = 15, FP = 5, FN = 5, and TN = 15.

V. CONCLUSION

In conclusion, after testing and comparing different machine learning techniques such as Random Forest, GMM, and GNB, we determined that SVM and KNN with $k = 5$ provided the greatest classification performance. The accuracy of the SVM model was over 82% and the accuracy of the KNN model with (k=5) was about 75%. These results suggest that SVM and KNN are the most suitable models for ethnic classification in spoken English by extracted features. As a result of the incorporation of MFCC, energy, zero-crossing rate, and pitch features, we were able to design an efficient ethnic group recognition system in speech. This goes to show that the type of methodology used in this project has been effective in delivering the goals of the project.

REFERENCES

[1] GeeksforGeeks, ”What is Speech Recognition?” [Online]. Available: https://www.geeksforgeeks.org/what-is-speech-recognition/. [Accessed: Jan. 16, 2025].

[2] ResearchGate, ”Approaches to Speaker Detection and Tracking in Conversational Speech.” [Online]. Available: https://www.researchgate.net/publication/220137320_Approaches_to_Speaker_Detection_and_Tracking_in_Conversational_Speech. [Accessed: Jan. 16, 2025].

[3] ResearchGate, ”Accent detection and speech recognition for Shanghai-accented Mandarin.” [Online]. Available: https://www.researchgate.net/publication/221479462_Accent_detection_and_speech_recognition_for_Shanghai-accented_Mandarin. [Accessed: Jan. 16, 2025].

[4] ResearchGate, ”Using Hybrid HMM/BN Acoustic Models: Design and Implementation Issues.” [Online]. Available: https://www.researchgate.net/publication/31157975_Using_Hybrid_HMMBN_Acoustic_Models_Design_and_Implementation_Issues. [Accessed: Jan. 16, 2025].

[5] ResearchGate, ”Support vector machines using GMM supervectors for speaker verification.” [Online]. Available: https://www.researchgate.net/publication/3343440_Support_vector_machines_using_GMM_supervectors_for_speaker_verification. [Accessed: Jan. 16, 2025].

[6] Nadia Pub, ”Automatic Identification of Vietnamese Dialects.” [Online]. Available: https://article.nadiapub.com/IJSIP/vol4_no2/8.pdf. [Accessed: Jan. 16, 2025].

[7] Academic Commons, ”Speech Accent Recognition using Machine Learning.” [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D8M61S68. [Accessed: Jan. 16, 2025].

[8] ResearchGate, ”Accent Recognition for Indian English using Acoustic Feature Approach.” [Online]. Available: https://www.researchgate.net/publication/258070241_Accent_Recognition_for_Indian_English_using_Acoustic_Feature_Approach. [Accessed: Jan. 16, 2025].

[9] ResearchGate, ”Automatic Identification of Vietnamese Dialects.” [Online]. Available: https://www.researchgate.net/publication/340231347_AUTOMATIC_IDENTIFICATION_OF_VIETNAMESE_DIALECTS. [Accessed: Jan. 16, 2025].

[10] ResearchGate, ”Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features.” [Online]. Available: https://www.researchgate.net/publication/307889236_Accent_Identification_by_Combining_Deep_Neural_Networks_and_Recurrent_Neural_Networks_Trained_on_Long_and_Short_Term_Features. [Accessed: Jan. 16, 2025].

[11] Scientific Research Publishing, ”Speech Emotion Recognition Based on Deep Learning Approaches,” Journal of Computer and Communications, vol. 8, no. 9, pp. 98-105, 2020. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=104256. [Accessed: Jan. 16, 2025].

[12] J. Hui, ”Speech Recognition Feature Extraction: MFCC, PLP,” Medium, 2020. [Online]. Available: https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9%7D. [Accessed: Jan. 16, 2025].

VI. Appendix

Code and results source:

https://drive.google.com/drive/folders/1BukxnELGAl-fFwdUfxHjeyulK4cZGvoW?usp=sharing