2020 11th International Conference on Electrical and Computer Engineering (ICECE)

A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification

Md. Fahad Hossain
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
fahad15-9600@diu.edu.bd

Md. Mehedi Hasan
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
mehedi15-9804@diu.edu.bd

Hasmot Ali
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
hasmot15-9632@diu.edu.bd

Md Rahmatul Kabir Rasel Sarker
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
raselsarker.cse@diu.edu.bd

Md. Toukirul Hassan
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
toukirul.cse0300.c@diu.edu.bd

2020 11th International Conference on Electrical and Computer Engineering (ICECE) | 978-1-6654-2254-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICECE51571.2020.9393038

***Abstract*—Speech is one of the primary modes of communication with a lot of identical features for measuring performance and behavior of human voice. Accent is an important element and can play a vital role in spoken language. In this paper, we propose a region detection approach of UK citizens by recognizing their accent from continuous speech. The ultimate goal of this paper is to detect the region of UK citizens from which region among Ireland, Midland, Northern England, Scotland, Southern England and Wales he/she belongs using continues speech. Firstly, we use Mel Frequency Cepstral Coefficient (MFCC) for extracting the feature from continuous speech. Then we applied several Machine Learning classifiers to train and test our model. After evaluating performance we find that k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Random Forest classifier provide comparatively better accuracy than others. We also perform a comparative analysis of these three algorithms. We got the best accuracy of 98.48% by applying k-NN classifier.**

Keywords*—Speech Processing, Speaker Recognition, Region Detection, Accent Classification

I. Introduction

Speech is a powerful medium for efficient communication. Having a bunch of unique features, speech is one of the most useful identifiers to recognize speakers. Each speech sound can be analyzed in terms of its phonetic features, the chunks of the sound that can each be autonomously controlled by the articulators. Accent is one of the powerful bases to differentiate speakers from the same language and different place. It is a unique way that groups of people who speak the same language sound and also an identifiable style of pronunciation. For successful communication, learning to quickly process accented speech is a prerequisite. The characteristic of accent is changing his perception in a human lifespan [1]. Interpersonal evaluations of speaker's accent is a useful technique to identify the local population and foreigners [2]. So, the accent dependent speaker recognition technique is one of the leading technologies in the field of speech recognition.

For extracting speech features there is a comparative study done by Tantisatirapong [3] for performing Thai Speech Recognition System depending on Accent using ESD, PSD, MFCC, and SPT algorithm. They conclude that the feature based on MFCC provides average accuracy of 93.81% for females and 89.34% for male voices. Chunrong [4] performs an approach that can perform speech feature extraction and detect the singular signal in a noisy environment.

Region detection is one of the important tasks of classifying people from a specific geographical area. It has a lot of applications including detecting the speaker from an unknown region, verifying the unknown region of crime suspect and speaker recognition. There are a lot of research is done for recognizing speakers from online meetings, television conversations, and live speech. Vinyals [5] performs a speaker recognition approach from an online meeting by using a single far-field microphone. A self-learning speech-controlled system for speaker identification and adaptation in terms of detecting unknown speaker is performed by Herbig [6-7] using Unsupervised Speech Controlled System. Amino [8] describe different factor that affects human speaker recognition application. They perform two different experiments to identify those effects. A speaker change detection in broadcast television is performed by Yin [9] where Bi-LSTM were used and said that Bi-LSTM method shows an improved result than conventional method. There is also a lot of speech-based speakers' gender, height, weight, age detection approach is done in different researchers [10-13].

The combination of speaker recognition for identifying different regions or other geographical identities is an important term in modern communication. A closely related work is done in different Indian languages to detect four different accents for Bengali, Gujarati, Malayalam and Marathi by Joseph [14]. They used Dynamic Time Warping algorithm to train the model, MFCC to extract features, and get 63.4% accuracy. The contribution is important for different languages, but it is quite difficult when someone has to work with the same language. Danao [15] perform a regional accent detection approach for Tagalog language in Philippines using several classifiers but Multi-Layer Perceptron (MLP) classifier did the best performance with

978-1-6654-2254-3/20/$31.00 ©2020 IEEE

93.33% accuracy. A study about determining American English and Indian English accented speech using Gaussian Mixture Modeling (GMM) is examined by Deshpande [16]. Mannepalli [17] introduces a method to identify three different accents namely Coastal Andhra, Rayalaseema and Telangana of Telugu language using Nearest Neighborhood Classifier and achieved 72% accuracy. For Chinese accent detection Long [18] perform a method based on RASTA - PLP algorithm extracting features known as short-time spectrum of each speech segment and record accuracy of 80.8% using Naïve Bayes classifier. Zheng [19] propose a new combination of accent discriminative acoustic features, accent detection, and an acoustic adaptation approach for accented Chinese speech recognition.

In this research, we have proposed a method that performs speaker identification from continues speech and recognizes the speaker on the basis of six different regions of the United Kingdom. We also perform a comparative analysis of three different algorithms used in this contribution.

II. METHODOLOGY

In this research, we follow the workflow showed in Fig.1

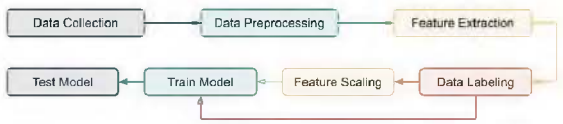

Fig.1: Overall Workflow

A. Dataset Discription

We used Crowdsourced high-quality UK and Ireland English Dialect speech data set from Demirsahin [20] for this work. This dataset is available on Google's open SLR. It contains audio data collected from six different regions of the United Kingdom. The total length of the recording is more than 31 hours. The duration of each audio is 6-7 seconds on average. There is a total of 17,877 audio samples where 10,627 audios were collected from males and 7,250 audios were collected from females. Southern England, almost 47% of the whole dataset.

Fig.1 shows the percentage of male-female ratio of total data.

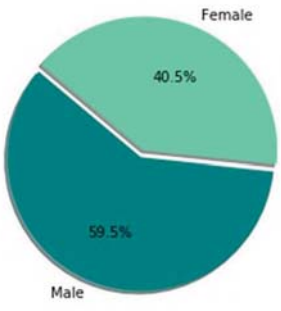

Fig.1 Ratio of male-female data

All the audio available in this dataset was recorded by the volunteers who identified themselves as native speakers of corresponding regions. Most of the data collected from Southern England, almost 47% of the whole dataset. Table. I shows how much data a region has.

TABLE I. NUMBER OF SAMPLE DATA FOR DIFFERENT REGION

Region	Sample
Ireland	450
Midland	696
Northern England	2847
Scotland	2543
Southern England	8492
Wales	2849

Fig.2 shows a comparison of the amount of data between the regions.

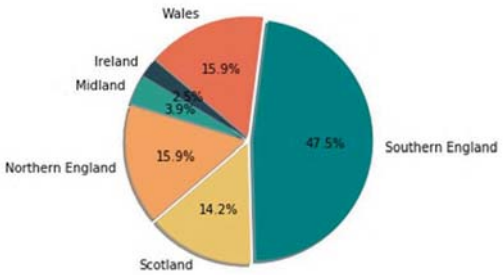

Fig.2: Percentage of data collected from the regions.

B. Feature Extraction

Before feature extraction, we did pre-processing of the audio file followed by Mehedi [21]. Audio features are the best way to represent an audio signal for machine learning algorithms. Among all the audio features, one of the most popular features are MFCC features. MFCC features are the short-time power spectrums of audio. It is calculated in some steps.

Algorithm: MFCC extraction

Input	:	Continuous audio speech
Output	:	Mel frequency cepstrum coefficients as features
1	:	Read continuous audio, A
2	:	Split the whole audio into some short frames
3	:	Calculate the power spectrum's periodogram estimate for each frame
4	:	Sum the energy of each filter from power spectrum by applying mel filter-bank
5	:	Calculate the logarithm of energies of each filter-bank.
6	:	Calculate the Discrete Cosine Transform (DCT) values of the logarithmic energies.
7	:	Take the first 20 DCT coefficients as features.

C. Data Labeling

Six regions are six types of accents. We labelled the accents by some integer numbers. Table. II shows accent names and corresponding integer values.

TABLE II. CORRESPONDING LABEL OF DIFFERENT REGION

Region Name	Label
Ireland	1
Midland	2
Northern England	3
Scotland	4
Southern England	5
Wales	6

D. Feature Salling

Distortion features bear a bad impact on the performance of the model. Feature scaling is needed to reduce the distortion of features. Here we do some simple transforms and compress the features into a short range. We tested the two most popular feature scaling techniques in this work.

1. Min-Max Scaling: Min-max Scaling is calculated from this equation [22]

$$x' = \frac{x - min(X)}{max(X) - min(X)} \quad .....(1)$$

2. Standard Scaling: Standard Scaling is calculated from this equation [22]

$$x' = \frac{x - \bar{x}}{\sigma} \quad .....(2)$$

III. EXPERIMENT SETUP AND RESULT

A. Experimental Setup

For training the model, we applied several machine learning algorithms. We applied these algorithms in two ways. In one way, we trained the model with the features extracted from the audio. In another way, we scaled the features and trained the model with scaled features. We observed that some algorithms work better with scaled features, while there is no impact of feature scaling on other algorithms. Among these algorithms, k-NN, SVM, and Random Forest performed better. Sample of the impact of feature scaling on different algorithms has been shown in table

TABLE III. DIFFERENT ALGORITHM AND THEIR ACCURACY

Algorithm	Accuracy		
	No scaling	Min-max scaling	Standard scaling
k-NN	90.67%	98.48%	98.23%
SVM	49.55%	63.15%	97.25%
Random Forest	93.47%	93.47%	93.47%

The confusion matrices of these three algorithms have been given in fig 3, 4, & 5 to measure the performance.

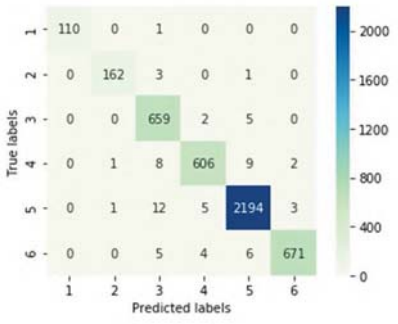

Fig. 3: Heat Map of k-Nearest Neighbors Confusion Matrix

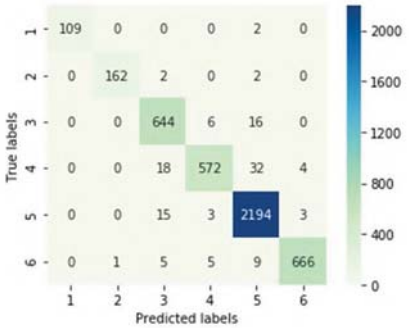

Fig. 4: Heat Map of Support Vector Machine Confusion Matrix

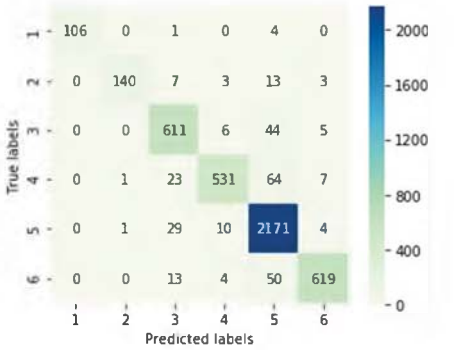

Fig.5: Heat Map of Random Forest Confusion Matrix

B. Result

We splitted the whole dataset into two segments of 75% for training the model and 25% for testing the model. For performance evaluation we have considered multiple parameters such as Precision, Recall, F1-score and Accuracy. The Label column represents corresponding region names given in Table II. The performance of the best three algorithms has been shown in tables IV, V & VI. In each table the parameter values of each label have been given from corresponding rows. The Macro-average values are the normal mean of each corresponding column. Support column has been used as weight for calculating the weighted average of each column. The accuracy of each label has been shown in corresponding rows at Accuracy column.

TABLE IV. PERFORMANCE OF K-NEAREST NEIGHBORS ALGORITHM

Label	Precision	Recall	F1-score	Support	Accuracy
1	1.00	0.99	1.00	111	0.99
2	0.99	0.98	0.98	166	0.98
3	0.96	0.99	0.97	666	0.99
4	0.98	0.97	0.98	626	0.97
5	0.99	0.99	0.99	2215	0.99
6	0.99	0.98	0.99	686	0.98
Macro-average	0.99	0.98	0.98	4470	0.98
Weighted average	0.98	0.98	0.98	4470	0.99

TABLE V. PERFORMANCE OF SUPPORT VACTOR MACHINE ALGORITHM

Label	Precision	Recall	F1-score	Support	Accuracy
1	1.00	0.98	0.99	111	0.98
2	0.99	0.98	0.98	166	0.98
3	0.94	0.97	0.95	666	0.97
4	0.98	0.91	0.94	626	0.91
5	0.97	0.99	0.98	2215	0.99
6	0.99	0.97	0.98	686	0.97
Macro-average	0.98	0.97	0.97	4470	0.97
Weighted average	0.97	0.97	0.97	4470	0.97

TABLE VI. PERFORMANCE OF RANDOM FOREST ALGORITHM

Label	Precision	Recall	F1-score	Support	Accuracy
1	1.00	0.95	0.98	111	0.95
2	0.99	0.84	0.91	166	0.84
3	0.89	0.92	0.91	666	0.92
4	0.96	0.85	0.90	626	0.85
5	0.93	0.98	0.95	2215	0.98
6	0.97	0.90	0.94	686	0.90
Macro-average	0.96	0.91	0.93	4470	0.91
Weighted average	0.94	0.93	0.93	4470	0.93

The overall Precision, Recall, F1-score values and final Accuracy of best three algorithms are shown in Table VII with corresponding Error Rate.

TABLE VII. PERFORMANCE IN DIFFERENT MODEL

Model	Precision	Recall	Fl-Score	Accuracy	Error Rate
k-NN	0.99	0.98	0.98	98%	0.02
SVM	0.98	0.97	0.97	97%	0.03
Random Forest	0.96	0,91	0.93	93%	0.07

IV. CONCLUSIONS

In this paper, we investigate the way of detecting speaker's region from six different regions of the United Kingdom using speech datasets named 'Crowdsourced high-quality UK and Ireland English Dialect speech data set' based on accent classification. We used MFCC for preprocessing the data and feature extraction. Then we perform several numbers of algorithm to measure the performance. But in this paper, we reported the performance of SVM, k-NN and Random Forest classifier to train and test our model. We also include the comparative analysis of these three different classifiers to introduce the best one for region detection.

Our findings demonstrate that k-NN provides the highest performance of 98.48% accuracy to detect speaker region where SVM and Random Forest provide an accuracy rate of 97.25% and 93.47%, respectively.

REFERENCES

[1] Cristia, Alejandrina et al. "Linguistic processing of accented speech across the lifespan." Frontiers in psychology vol. 3 479. 8 Nov. 2012, doi:10.3389/fpsyg.2012.00479J.

[2] Fuertes, Jairo & Potere, Jodi & Ramirez, Karen. (2002). Effects of speech accents on interpersonal evaluations: Implications for counseling practice and research. Cultural diversity & ethnic minority psychology. 8. 346-56. 10.1037/1099-9809.8.4.347.

[3] Tantisatirapong, Suchada & Prasoproek, Chalisa & Phothisonothai, Montri. (2018). Comparison of Feature Extraction for Accent Dependent Thai Speech Recognition System. 322-325. 10.1109/CCE.2018.8465705.

[4] X. Chunrong, Z. Jianhuan and L. Fei, "A Dynamic Feature Extraction Based on Wavelet Transform for Speaker Recognition," 2007 8th International Conference on Electronic Measurement and Instruments, Xi'an, 2007, pp. 1-595-1-598, doi: 10.1109/ICEMI.2007.4350520.

[5] Vinyals, Oriol & Friedland, Gerald. (2008). LIVE SPEAKER IDENTIFICATION IN MEETINGS - "WHO IS SPEAKING NOW?.

[6] Herbig, Tobias & Gerl, Franz & Minker, Wolfgang. (2010). Detection of Unknown Speakers in an Unsupervised Speech Controlled System. 6392. 25-35. 10.1007/978-3-642-16202-2_3.

[7] Herbig, Tobias & Gerl, Franz & Minker, Wolfgang. (2012). Self-learning speaker identification for enhanced speech recognition. Computer Speech & Language. 26. 210-227. 10.1016/j.csl.2011.11.002.

[8] Amino, Kanae & Arai, Takayuki. (2009). Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. Acoustical Science and Technology. 30. 89-99. 10.1250/ast.30.89.

[9] Ruiqing Yin, Hervé Bredin, Claude Barras. Speaker Change Detection in Broadcast TV Using Bidi-rectional Long Short-Term Memory Networks. Interspeech 2017, Aug 2017, Stockholm, Sweden. 10.21437/ Interspeech.2017-65. hal-01690244

[10] Badhon, S M & Rahaman, Md & Rupon, Farea. (2019). A Machine Learning Approach to Automating Bengali Voice Based Gender Classification. 55-61. 10.1109/SMART46866.2019.9117385.

[11] Bahari, Mohamad & Van hamme, Hugo. (2011). Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization. 1 - 6. 10.1109/BIOMS.2011.6052385.

[12] Hansen, John & Williams, Keri & Boril, Hynek. (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. The Journal of the Acoustical Society of America. 138. 1052. 10.1121/1.4927554.

[13] Mporas, I., Ganchev, T. Estimation of unknown speaker's height from speech. Int J Speech Technol 12, 149–160 (2009). https://doi.org/10.1007/s10772-010-9064-2

[14] J. Joseph and S. S. Upadhya, "Indian accent detection using dynamic time warping," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2814-2817, doi: 10.1109/ICPCSI.2017.8392233.

[15] G. Danao, J. Torres, J. V. Tubio and L. Vea, "Tagalog regional accent classification in the Philippines," 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Manila, 2017, pp. 1-6, doi: 10.1109/HNICEM.2017.8269545.

[16] S. Deshpande, S. Chikkerur and V. Govindaraju, "Accent classification in speech," Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), Buffalo, NY, USA, 2005, pp. 139-143, doi: 10.1109/AUTOID.2005.10.

[17] K. Mannepalli, P. N. Sastry and V. Rajesh, "Accent detection of Telugu speech using prosodic and formant features," 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, 2015, pp. 318-322, doi: 10.1109/SPACES.2015.7058274.

[18] Long, Zhang & Yunxue, Zhao & Peng, Zhang & Ke, Yan & Wei, Zhang. (2015). Chinese accent detection research based on RASTA - PLP algorithm. 31-34. 10.1109/ICAIOT.2015.7111531.

[19] Zheng, Yanli & Sproat, Richard & Gu, Liang & Shafran, Izhak & Zhou, Haolang & Su, Yi & Jurafsky, Daniel & Starr, Rebecca & Yoon, Su-Youn. (2005). Accent detection and speech recognition for Shanghai-accented Mandarin.. 217-220.

[20] Demirsahin, Isin & Kjartansson, Oddur & Gutkin, Alexander & Rivera, Clara. (2020). Opensource Multispeaker Corpora of the English Accents in the British Isles

[21] M. M. Hasan, H. Ali, M. F. Hossain and S. Abujar, "Preprocessing of Continuous Bengali Speech for Feature Extraction," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225469.

[22] Wikipedia link: [online], https://en.wikipedia.org/wiki/Feature_scaling