Speaker Region Identification

Aurosweta Mahapatra, Naveen Vikraman, Swagath Babu

University of California, Los Angeles, USA

aurosweta99@ucla.edu, naveenvikraman@g.ucla.edu , swagathb18@g.ucla.edu

Abstract

In this paper we will be discussing different approaches for "Speaker Region Identification" using Corpus of Regional African American Language (CORAAL) dataset. A speaker identification system that predicts the city of origin of the speaker of a given utterance can be useful in several applications, such as forensic investigations, speech recognition systems, and language and dialect research. By identifying the city of origin of a speaker, we can better understand the linguistic and cultural differences that exist across different regions. We faced some difficulty categorizing the speaker's region from the give dataset as the speaker identification system will depend on several factors, such as the quality and size of the dataset, the selection of relevant features, and the choice of classification algorithm. Additionally, the system may be affected by factors such as the speaker's age, gender, and education level, which could introduce variability in the speech signal. We used multiple feature extraction techniques to overcome these challenges.

1. Introduction

We have implemented a speaker identification system that predicts the city of origin of the speaker of a given utterance. The dataset used for this project is Corpus of Regional African American Language (CORAAL). The dataset contains speakers each belonging to one of five different US cities: 1) Rochester, NY (ROC), 2) Lower East Side, Manhattan, NY (LES), 3) Washington DC (DCB), 4) Princeville, NC (PRV), or 5) Valdosta, GA (VLD). The utterances in the corpus with length greater than 10 seconds are only selected. In addition, a few utterances have been corrupted by a 10dB babble noise masker. The babble noise is linear in all the files. Apart from this given dataset, the model is also be tested on a blind test set. The blind test set consists of a different set of speakers from the above cities. We have used different feature extraction techniques with libraries like Librosa and Speechpy. We used MFCC(13), MFCC(39), LPC, Energy Normalization techniques, Mean and Variance Normalization, Discrete Wavelet Transform, Noise reduction using Specrtral Gating, Zero crossing, Root Mean Square Error, Log bank and Chroma.

2. Background

One of the most interesting applications of speech processing is Speaker Region Identification. It has a vast implementation in the present world, but due to the complexity of algorithms, limited dataset availability, and many other speech-related factors, it's accuracy in performance and use is affected. In this project, we have implemented various feature extraction techniques with different libraries and their combinations to boost the accuracy with and without noise for the CORAAL dataset.

2.1. Dataset

We have implemented this project with the use of CORAAL dataset. The CORAAL dataset (Corpus of Regional African American Language) is a publicly available dataset of spoken language samples from African American speakers across the United States. It was developed to provide a representative sample of the diversity of African American English (AAE) dialects and regional variations. The dataset contains speakers each belonging to one of five different US cities: 1) Rochester, NY (ROC), 2) Lower East Side, Manhattan, NY (LES), 3) Washington DC (DCB), 4) Princeville, NC (PRV), or 5) Valdosta, GA (VLD). The utterances in the corpus with length greater than 10 seconds are only selected. In addition, a few utterances have been corrupted by a 10dB babble noise masker. The babble noise is linear in all the files. The sampling frequency of the wav files provided in the dataset is 44.1kHz. The recordings in the CORAAL dataset were collected using a variety of methods, including interviews, group discussions, and monologues. The speech samples were transcribed and annotated using the International Phonetic Alphabet (IPA), allowing researchers to analyze the phonetic, grammatical, and lexical features of the AAE dialects spoken by the participants.

There is 4372 training files , 447 test clean files and 347 test noise files. It is an imbalanced dataset with more samples from ROC region as illustrated below.

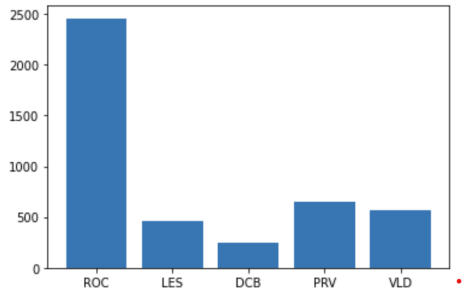

Figure 1: *: Barplot of distribution of training labels*

2.2. Application

Forensic investigations: Speaker region identification can be used in forensic investigations to identify the origin of anony-

mous or unknown speakers, such as in criminal cases or terrorist threats. This information can help law enforcement agencies narrow down their search for potential suspects or gather additional evidence. Speech recognition systems: Speaker region identification can also be used in speech recognition systems to improve their accuracy. By identifying the geographic origin of the speaker, the system can adapt its recognition algorithms to better match the speaker's dialect and accent, which can reduce errors and improve the overall user experience. Language and dialect research: Speaker region identification can also be used in language and dialect research to better understand the linguistic and cultural differences that exist across different regions. By analyzing the speech patterns of speakers from different regions, researchers can identify unique features and characteristics that are specific to particular dialects or accents. This can help improve language education and promote cultural diversity.

3. Project Description

We have tried various combinations of features and the summary of the results is given below.

In this paper, we will be focusing on three feature extraction

Test Noisy Acc	Test Clean Acc	Noise Reduce	MFCC (13)	MFCC (39)	Log Bank (26)	DWT MFCC (39)	Pitch MFCC (13)	MFCC (8)	ZCR	CMV WN	RMSE	MFCC (13) speec hpy
61.6	77.6		1									
61.6	78.9			1								
63.4	71.3				1							
50.14	88.8					1						
65.9	65.3			1	1							
66.8	72.4			1	1	1	1			1		
69.7	59.2	1	1									
69.7	59.7	1		1								
66.2	68.9	1			1							
76.9	64.6	1		1	1							
52.4	82.3							1				
63.4	78.2							1	1			
63.9	78.7							1			1	
75.8	81.2				1			1	1		1	
63.1	63.7											1

Figure 2: *: Combination of Feature extraction methods. One's in the table represents the corresponding feature in the column used for the particular accuracy value. The highlighted row represents our three best outcomes. Best test clean, best test noisy and best balance accuracy values for the given dataset.*

methods for which we got the best accuracy value compared to all other methods. These methods are based on 1) the best accuracy value for test noisy, 2) the best accuracy value for test clean, and 3) balanced accuracy values for both test clean and test noisy. Furthermore, these models were tested on a hidden dataset and the accuracy achieved for that will be discussed in this section

3.1. Features and Methods

We have used different feature extraction techniques with libraries like Librosa and Speechpy. The techniques we have used are Mel Frequency Cepstrum Coefficient(MFCC(13)),MFCC(39MFCC(13)+Delta(13)+Double Delta(13)), MFCC(8), Noise Reduction, Log Bank(26), Discrete wavelet Transform, Root mean square error, Pitch MFCC, Cepstral Mean Variance Window Normalization, Zero Crossing. In this paper we will be discussing three feature extraction methods for which we got the best accuracy value compared to all other methods.

3.1.1. Best Test Clean Model

After trying out multiple feature extraction model, we got the best result for test clean using Discrete Wavelet Transform followed by MFCC(39) features[4]. Discrete Wavelet Transform (DWT) [1] is a useful feature extraction technique in speech processing. Wavelet analysis allows for the decomposition of a signal into a set of wavelet coefficients, which can capture both the frequency and temporal characteristics of the signal. This is important because accents can be characterized by both the pitch and duration of certain phonemes. Additionally, wavelet analysis can capture high-frequency components of the signal that are often lost in traditional spectral analysis methods such as Fourier analysis or Mel frequency cepstral coefficients (MFCCs). Delta coefficients represent the rate of change of the MFCCs over time and are computed by taking the first-order temporal derivative of the MFCCs. Basically, it is the slope of the MFCC curve over time and can capture the speech signal's speed of articulation. Double delta coefficients represent the acceleration of the MFCCs over time and are computed by taking the second-order temporal derivative of the MFCCs. They can be thought of as the curvature of the MFCC curve over time and can capture the fine-grained details of the speech signal's temporal dynamics. We managed to get an accuracy of 88.8% for the test clean dataset but the accuracy for test noisy for this model is 50.14%. The reason for this low accuracy for test noisy dataset is that we did not perform any noise reduction for this model. It can also be observed that the identification for LES region is poorest of all the five regions. The confusion matrix in figure 3 and classification report in figure 4 represents the visual and mathematical representation for the same.

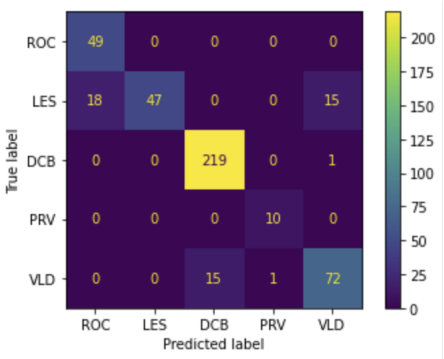

Figure 3: *: Confusion matrix for the test clean dataset using DWT+MFCC(39) feature extraction technique.*

	precision	recall	f1-score	support
ROC	0.73	1.00	0.84	49
LES	1.00	0.59	0.74	80
DCB	0.94	1.00	0.96	220
PRV	0.91	1.00	0.95	10
VLD	0.82	0.82	0.82	88
accuracy			0.89	447
macro avg	0.88	0.88	0.86	447
weighted avg	0.90	0.89	0.88	447

Figure 4: *: Classification report for the test clean dataset using DWT+MFCC(39) feature extraction technique.*

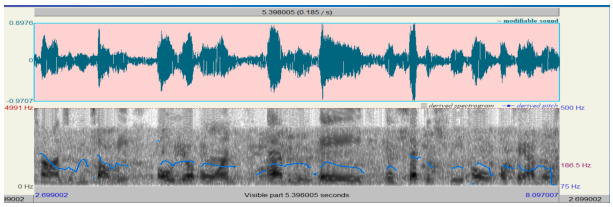

Figure 5: *The spectrogram of a noisy audio signal*

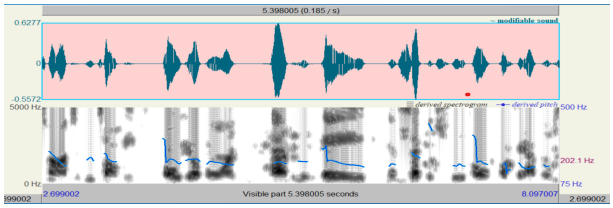

Figure 6: *The spectrogram of noise reduced audio signal*

3.1.2. Best Test Noise Model

We thought of reducing the noise in the data before feeding it into the classification algorithm. We chose Non-statoinary noise reduction using Spectral gating as the noise reduction method. We chose a non-stationary method as babble noise is non-stationary. Non-stationary means that the noise signal statistics such as mean, variance etc changes over time which happens when people are talking the background. Overview of steps in Noise reduction using Spectral gating:

- A spectrogram is calculated over the noise audio clip.
- A time-smoothed version of the spectrogram is computed using an IIR filter aplied forward and backward on each frequency channel.
- Statistics are calculated over spectrogram of the the noise in frequency domain
- The mask is smoothed with a filter over frequency and time
- The mask is applied to the spectrogram of the signal, and is inverted

From Figure 5 and Figure 6, we can see that the noise reduction techniques has been able to remove a lot of babble noise frequency content. While playing the reconstructed noise reduced signal, it was observed that the spectral gating indeed helped in noise reduction. But the stops after each word are not that clear.

We applied Noise reduction on both training and test data before being passed through the classification algorithm. A combination of MFCC with delta and double delta(39) and Log-bank filter energy was used as the features for the model having best accuracy on noisy test data.

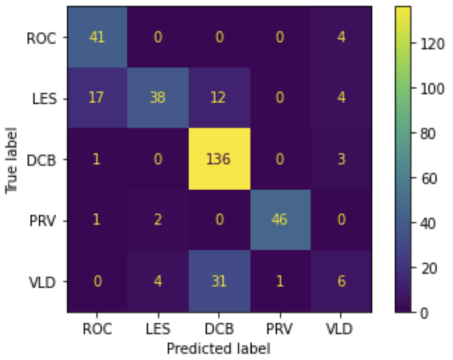

Figure 7: *: Confusion matrix for the test noisy dataset using noisereduce and MFCC(39) and logbank(26) feature extraction technique.*

	precision	recall	f1-score	support
ROC	0.68	0.91	0.78	45
LES	0.86	0.54	0.66	71
DCB	0.76	0.97	0.85	140
PRV	0.98	0.94	0.96	49
VLD	0.35	0.14	0.20	42
accuracy			0.77	347
macro avg	0.73	0.70	0.69	347
weighted avg	0.75	0.77	0.74	347

Figure 8: *: Classification report for the test noisy dataset using noisereduce and MFCC(39) and logbank(26) feature extraction technique.*

From the classification report we can see that regions VLD and LES is hard to classify. We have also tried using pitch feature in concatenation with existing features, but it didnt give any performance improvement. We think it is because logbank energy feature inherently captures information about pitch and changes in pitch.

3.1.3. Balanced Model for Test Clean & Test Noise

In this model, the goal was to increase classification accuracy for the same features, when we performed noise removal in the previously discussed model in (3.1.2), we were able to improve the noise accuracy from the baseline model but in some cases, we were facing issue in classifying clean data, as some of the information was lost during noisy removal which did not affect too much but it is definitely considered when it comes to classification on the entire dataset.

Here we used a combination of different feature extraction methods such as mfcc (13), ZCR, RMSE, and log filter bank, the motivation to use ZCR and RMSE is robust to noisy data [2] and commonly used in speech detection, noise reduction, and speaker identification. ZCR captures the rate of change of speech sign which gives information about the frequency components of the signal. In simple words, it is a measure of the number of times the signal crosses the zero axis. RMSE is help-

ful in capturing the intensity of the audio or loudness of the audio, which provides information about phonemes and syllables which helps in detecting and reducing the noise component, but it is not explicitly used for noise reduction. Overall RMSE gives us the measure of the average energy of the speech signal.

We want to incorporate these methods along with the log bank

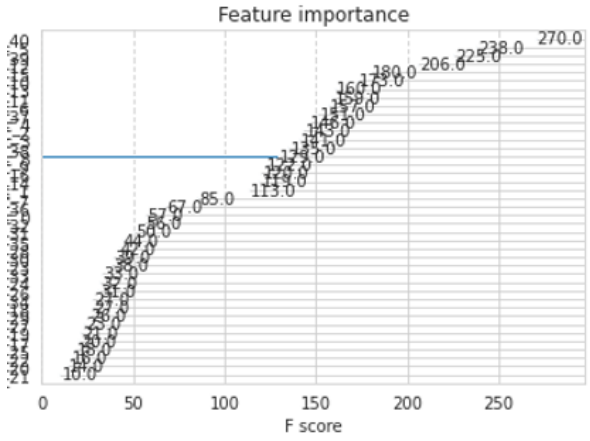

Figure 9: *: Feature Importance F1 score.*

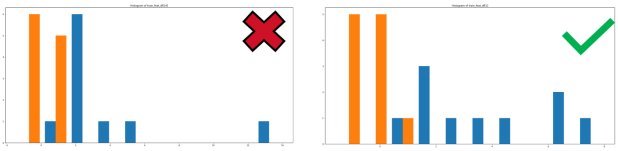

Figure 10: *: Histogram (feature selected based on good spread of data with less empty bars).*

(26) which helped us in the previous model to improve noisy accuracy as this filter inherits the triangular band pass filter in Mel-Scale frequency which maps to human hearing and thus helps to capture features even in noisy situations. In one of our tests by using only mfcc with some important feature which was selected based on the F1 score-feature important (figure 9), and histogram plot (figure 10) [3], we were able to achieve 82% accuracy in clean data hence we thought by selecting important features from mfcc, ZCR, RMSE, and log bank could beat the baseline model, as a result, we were able to surpass the noisy accuracy of baseline but not the clean data.

Hence, we tried different feature selections like permutation feature (figure 11) selection along with shap which was helpful to some extent. We tried to understand the correlation between every feature and based on the observation on the correlation matrix (figure 11) of all the 41 features that is been extracted and tested for highly correlated features which resulted in good noisy accuracy but poor clean accuracy hence selected the features from permutation feature selection method in which most of the features are not highly correlated which we used it for training and testing our model, which improved the accuracy to 72% for clean and 70% for noisy data.

From confusion matrix (figure 13), it can be observed that LES and VLD classification is improved from (figure 3). This can be due to using all 39 features from mfcc in which first 13 are actual mfcc and next sets of 13 are their first order and sec-

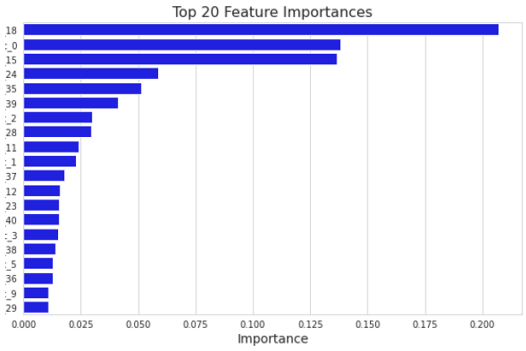

Figure 11: *: Permutation feature importance plot.*

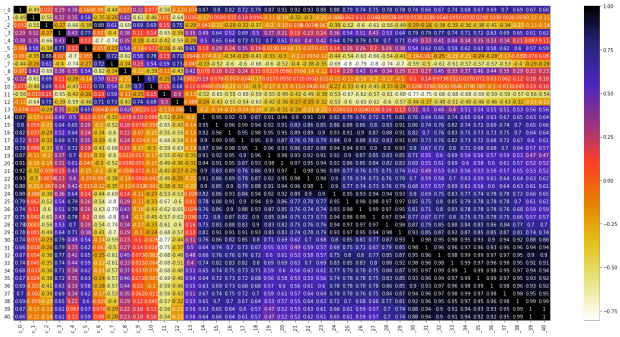

Figure 12: *:Correlation Matrix(dark region-¿high correlation)*

ond order derivatives, which can result in correlation in simple words they are velocity and acceleration of the signal which can make the classifier more challenging in classifying some of the speech with high correlation. From (figure 14) the classification report shows F1 score of VLD is less when compared to other regions which makes it hard to classify, but we were able to improve LES classification with good recall score and F1 score using mentioned feature extraction method.

According to the classification report from (figure 16), the F1 scores for the regions VLD and LES in clean data improved, with the F1 score for VLD rising by 180% and the F1 score for LES falling by 40%. Nonetheless, the model outperformed the prior model(3.1.2) by correctly classifying at least 50% of the test data. precion score increase by 108% in VLD from best noise model and 220% increase in VLD recall score for noise test data which is pretty good in classification for both dataset.

When we tried to analyze the pitch of VLD and LES data we observed that it had a lot of stops which makes it harder for the classifier to make the decision. These stops can be filler words that can represent the cognitive or emotional state of the speaker which is hard for the machine to understand and classify. However, this can be overcome by removing the filler or stop words but this will not be a universal approach in classifications.

3.1.4. Hidden Dataset Performance

After testing the performance of our models on the given dataset, we checked it's performance for a hidden dataset. Our best test clean model which uses DWT+MFCC(39) feature extraction techniques performed very well for the hidden dataset with accuracy of 90.02

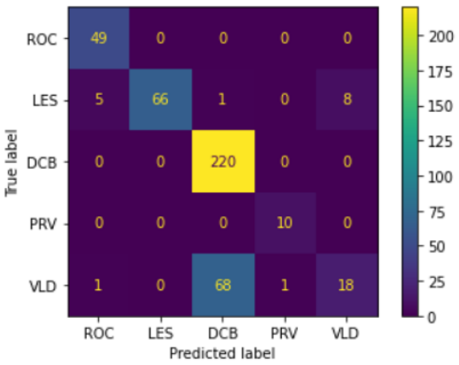

Figure 13: *: Confusion matrix for the test clean dataset using MFCC+ ZCR+ RMSE+log bank feature extraction technique.*

	precision	recall	f1-score	support
ROC	0.89	1.00	0.94	49
LES	1.00	0.82	0.90	80
DCB	0.76	1.00	0.86	220
PRV	0.91	1.00	0.95	10
VLD	0.69	0.20	0.32	88
accuracy			0.81	447
macro avg	0.85	0.81	0.80	447
weighted avg	0.81	0.81	0.77	447

Figure 14: *: Classification report for the test clean dataset using MFCC+ ZCR+ RMSE+log bank feature extraction technique.*

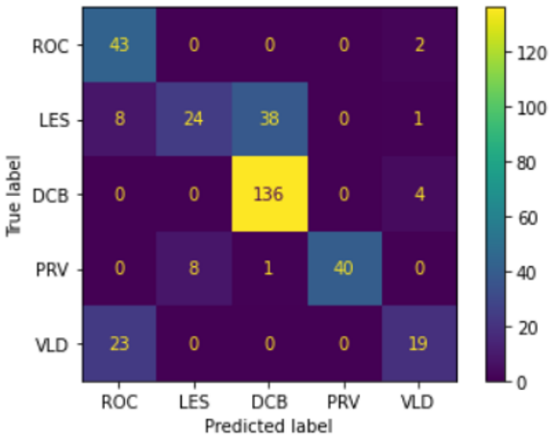

Figure 15: *: Confusion matrix for the test clean dataset using MFCC+ ZCR+ RMSE+log bank feature extraction technique.*

4. Conclusion and future directions

In this study, we looked into the XGboost classifier that is used to identify speakers based on their geographic location. We looked into employing several feature extraction techniques separately as well as combining numerous features to increase the model's accuracy for both clean and noisy datasets. When the best test clean model was tested with a hidden test dataset, the clean accuracy reached 90%, and the noisy accuracy was 84%. As a consequence, we produced three models: the best test clean model, the best test noisy model, and the balanced model. This demonstrates that the model is capable of classifying with good accuracy, which is a 13% improvement for clean data and a 42% improvement over the baseline model's noisy data.

	precision	recall	f1-score	support
ROC	0.58	0.96	0.72	45
LES	0.75	0.34	0.47	71
DCB	0.78	0.97	0.86	140
PRV	1.00	0.82	0.90	49
VLD	0.73	0.45	0.56	42
accuracy			0.76	347
macro avg	0.77	0.71	0.70	347
weighted avg	0.77	0.76	0.73	347

Figure 16: *: Classification report for the test clean dataset using MFCC+ ZCR+ RMSE+log bank feature extraction technique.*

In our future study, more focus should be placed on speech features that are more useful. Also, a deeper learning model that is more sophisticated can generalize more and enhance the accuracy.

5. References

[1] Deshmukh, Ratnadeep & Ghule, Kishori. (2015). Feature Extraction Techniques for Speech Recognition: A Review. International Journal of Scientific and Engineering Research. 6. 143-147.

[2] Chauhan, N., Isshiki, T., & Li, D. (2019). Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). IEEE. https://doi.org/10.1109/ccoms.2019.8821751

[3] Ghosh, M., Guha, R., Singh, P. K., Bhateja, V., & Sarkar, R. (2019). A histogram based fuzzy ensemble technique for feature selection. In Evolutionary Intelligence (Vol. 12, Issue 4, pp. 713–724). Springer Science and Business Media LLC. https://doi.org/10.1007/s12065-019-00279-6

[4] Jain, K., Chaturvedi, A., Dua, J., & Bhukya, R. K. (2022). Investigation Using MLP-SVM-PCA Classifiers on Speech Emotion Recognition. In 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE. https://doi.org/10.1109/upcon56432.2022.9986457