National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

# Automatic Dialect Classification using SVM

**Achala H A, Avni Sharma, Rakshitha G K, Ramya V, Ramesh G**
Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, Karnataka, India

## ABSTRACT

Automatic Dialect Classification has attracted researchers in the field of speech signal processing. Dialect is defined as the language characteristics of a specific community. As such, dialect can be recognized by speaker phonemes, pronunciation, and traits such as tonality, loudness, and nasality. Dialect classification is a substantial tool in speech recognition and has the potential to improve the efficiency of Automatic Speech Recognition systems. This paper presents a study of different dialects in English language (American) and features that are useful for their classification. The experiment demonstrates that there are several features of the speech signal which are conducive for recognizing different dialects within a language such as chroma features and spectral features. Other speech features including MFCC and FDLP were also used with these features in order to improve the performance of the classifier. The supervised machine learning classifier that has been used in our research is the Support Vector Machine. Some refinements were introduced to the existing chroma feature extraction processes to make them more suitable for speech signal classification.

**Keywords:** Dialect classification, MATLAB R2014a, chroma features, spectral features, Support Vector Machine, MFCC.

## I. INTRODUCTION

Dialect classification[9] is a substantial tool in speech recognition and has the potential to improve the efficiency of Automatic speech Recognition systems. In this study we employ the definition of dialect as a pattern of pronunciation and/or vocabulary of languages used by a community of native speakers belonging to the same geographical region. Due to such differences in dialects the same language has multiple versions across different regions around the globe. Dialect classification also plays a key role in characterizing speaker traits and knowledge estimation, which can then be utilized to build dynamic lexicons by selecting alternative pronunciations and generate pronunciation modelling via dialect adaptation. In this project we plan to study about different dialects in American English language and features that are useful for their classification. An experiment was conducted to demonstrate that there are several features of the speech signal which are conducive for recognizing different dialects within a language such as Chroma features[2] and spectral features[3], etc. Other speech features including MFCC[5] and FDLP[4] can also be used with these features in order to improve the performance of the classifier. Chroma features[2] can be primarily used to classify music signals into different genre of music but the process of separating frequencies into bins is also applicable for classifying speech signals. Some refinements can be introduced to the existing Chroma feature extraction processes to make them more suitable for speech signal classification.

### A. Datasets

The first dataset used is a text dependent dataset which consists of total 9 dialect classes with 67 speech samples in each class. The second dataset is spontaneous or text independent dataset and contains 9 dialect classes with 72 speech samples in each class. The third dataset is ''TIMIT dataset" and has 8 dialect classes from different regions of America. In this dataset the number of speech signal samples varies from one dialect class to another. TIMIT dataset is also text dependent but the variations between the dialects of different classes are very circumstantial and difficult to observe.

TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. Each transcribed element has been delineated in time. The corpus contains a total of 6300 sentences, 10 sentences spoken by 630 speakers selected from 8 major dialect regions of the USA. 70% of the speakers are male, 30% are female. The text corpus design was done by the Massachusetts Institute of Technology (MIT), Stanford Research Institute and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and has been maintained, verified and prepared for CDROM production by the American National Institute of Standards and Technology (NIST)

The dialect regions are:

dr1: New England
dr2: Northern
dr3: North Midland
dr4: South Midland
dr5: Southern
dr6: New York City
dr7: Western
dr8: Army Brat (moved around)

In our experiment we have ignored the dr1 since it belongs more to a British dialect.

## II. METHODOLOGY

In the given Figure-1 input signal is from the datasets which has been explained in section II. The data is cleansed[1] and then the feature extraction phase starts where several features are extracted namely chroma features[2], MFCC/FDLP[5,4] features and spectral features[3], also named as other features here. Further proceeding in the experiment is the training and testing phase with the help of the SVM classifier[6,7]. The dialects are then suitably classified according to their dialects and the accuracy of the system is noted. The implementation has been done using MATLAB[10].

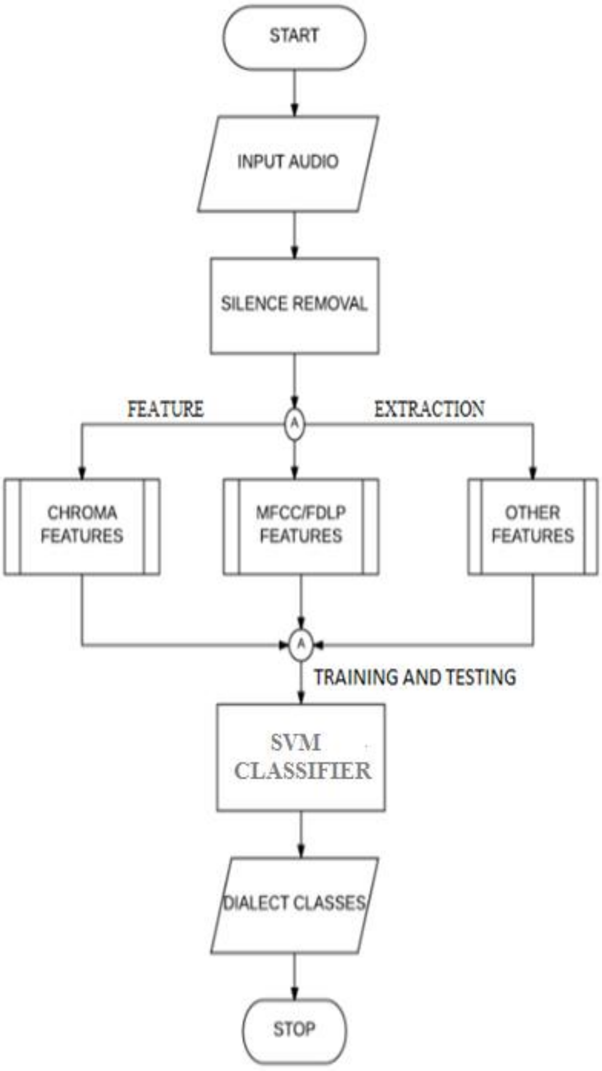

**Figure 1.** Flow of control of system

## III. FEATURE EXTRACTION

### A. Chroma Features

The primary aim of this experiment was to employ chroma features in our process of dialect classification. These features were originally built to classify music samples into different genre of music[2]. Due to its capability to use frequency bands for classification of signals it was quite practical to use them for dialect classification as well. People belonging to different regions intuitively use different proportions of these frequency bands in their regular speech. Therefore it is viable to distinguish between speech signals of different dialects on considerate observation of these frequency bands.

A total of 12 chroma features were extracted corresponding to the normalized energy of each of the frequency bins. It was later confirmed that only 10 such bins were enough for speech signal classification as the remaining 2 bins always remained unused for all the speech samples taken into consideration.The speech signal is first segmented into a number of frames and then the chroma features are extracted from all these frames.
After applying these changes it was found that the efficiency of classification process improved significantly in comparison to the previous version of the same features. A total of 20 features were extracted from the speech samples.

### B. Other Features

Other than the chroma features, spectral features[3] were also involved in the classification process. These are obtained by converting the time based signal into energy domain using the Fourier transform. It includes energy entropy, spectral centroid, spectral entropy, spectral flux, spectral roll-off and harmonic features. These were used with chroma features in order to improve classification accuracy of the speech signal. A total of 16 such features were extracted from the speech samples.

### C. MFCC Features

MFCC (Mel Frequency Cepstral Coefficients) features[5] are widely used in speech recognition process. MFCC are used because it is designed using the knowledge of human auditory system and is used in every state of speech recognition system. It is a standard method for feature extraction in speech recognition tasks.

They inherently have only 13 features but the deltas and delta-deltas which are also known as differential and acceleration coefficients are also extracted from the speech signal. The presence of these 26 extra features contributed from deltas and delta-deltas improve the performance of MFCC features. The MFCC vector describes only the power spectral envelope of a single frame, but speech also has information in the dynamics like the trajectories of the MFCC coefficients over time. Therefore it was found more profitable to calculate the MFCC trajectories and append them to original features. A total of 78 features were extracted from the speech samples.

### D. FDLP Feature

FDLP(Frequency Domain Linear Prediction) features[4] have three different types. The first is FDLP-s features which are quite similar to MFCC features. These are alternatively used and compared against MFCC features throughout the experiment to find out which one of them is more suitable for dialect classification process. The second type of feature in FDLP is FDLP-m which are long term modulation features. The third variety in FDLP is FDLP-PLP features which are short term features resembling two PLP features. In our experiment, we have use FDLP-s features for the classification process. For FDLP-s a total of 78 features were extracted from speech samples.

## IV. EXPERIMENTS AND RESULTS

A SVM model[6,7] was used for classification of speech signals into various dialects. Each of the above mentioned datasets were divided into five equal parts and a 5 fold cross validation method was applied on them. Thus using 80% of data for training and remaining 20% of data for testing.

### A. Selecting appropriate features

Initially a number of experiments were conducted on text dependent dataset over the choice of optimum features for dialect classification. The supervised machine learning classifier is the Support Vector Machine[6,7]. Experiments were done for folds=5 and folds =10. The results that are going to be discussed are for folds=5.

The first comparison was made to elect a set of chroma features from various chroma features and its derivatives available. This includes chroma features, CENS (chroma energy normalized statistics) and CRP (chroma DCT- reduced log pitch). The result of the experiment were as follows:

**Table 1.** Comparison between various chroma features for dataset 1(Text Dependent)

| Feature Name | No. of Features | Accuracy(%) |
|---|---|---|
| CENS | 24 | 64.667 |
| CRP | 24 | 45.21 |
| Chroma | 24 | 85.271 |

**Table 2.** Comparison between various chroma features for dataset 2(Text Independent)

| Feature Name | No. of Features | Accuracy(%) |
|---|---|---|
| CENS | 24 | 58.139 |
| CRP | 24 | 45.052 |
| Chroma | 24 | 69.767 |

The next comparison was between original chroma features and the modified chroma features to decide which of them was more suitable for speech signal classification. In the modified chroma feature extraction process the features corresponding to 4th and 7th class were removed as none of the frequency bands were assigned to those bins for any speech sample. The comparison between them can be observed from the following result:

**Table 3.** Comparison between original chroma features and updated features for dataset 1(Text Dependent)

| **Feature Name** | **No. of Features** | **Accuracy(%)** |
|---|---|---|
| Chroma Original | 24 | 85.271 |
| Chroma Updated | 20 | 95.16 |

**Table 4**

| Feature Name | No. of Features | Accuracy(%) |
|---|---|---|
| Chroma Original | 24 | 69.767 |
| Chroma Updated | 20 | 91.318 |

Table 4 Comparison between original chroma features and updated features for dataset 2(Text Independent) In the next comparison it is observed that text dependent dataset when MFCC features[5] combined with chroma[2] and spectral features[3] were more accurate than the FDLP-s features combined with chroma and spectral features whereas for the text independent[8] dataset MFCC features combined with chroma and spectral features were equally accurate as the FDLP-s features combined with chroma and spectral features. The accuracy of classification of speech signals into their dialect classes were observed as follows:

**Table 5.** Comparison between MFCC and FDLP-s features for dataset 1(Text dependent)

| Feature Name | No. of Features | Accuracy(%) |
|---|---|---|
| MFCC(13) | 26 | 37.818 |
| MFCC(rasta) | 78 | 67.5 |
| FDLP-s | 78 | 90 |
| MFCC +chroma+Others | 114 | 96.39 |
| FDLP-s+ Chroma+Others | 114 | 90.31 |

**Table 6.** Comparison between MFCC and FDLP-s features for dataset 2(Text Independent)

| Feature Name | No. of Features | Accuracy(%) |
|---|---|---|
| MFCC(13) | 26 | 84.03 |
| MFCC(rasta) | 78 | 95.8 |
| FDLP-s | 78 | 80.63 |
| MFCC +chroma+Others | 114 | 87.115 |
| FDLP-s+ Chroma+Others | 114 | 86.656 |

So our feature vector contained a total of 114 features including 78 MFCC or 78 FDLP-s features depending upon the dataset which the classification was taking place along with 20 chroma features and 16 other features.

The third dataset that is "TIMIT" dataset was also used for the classification process but since the dialect classes in that dataset were quite similar to each other the accuracy of classification of the speech samples were not impressive. The dataset consisted of different dialects from various regions of America. The following were the observed results with the selected feature sets:

**Table 7.** Comparison between various features for TIMIT dataset

| Feature Name | Accuracy |
|---|---|
| MFCC(13) | 19.7 |
| MFCC(rasta) | 22.9 |
| Chroma(Original) | 21.687 |
| Chroma(Updated) | 20.9 |
| MFCC+Chroma+Others | 20.482 |
| FDLP-s | 18.765 |
| FDLP-s+Chroma+Others | 21.205 |

In the initial stages of the experiment the number of folds being used was ten with the same number of features as mentioned in section (). But due to it's over-fitting behaviour it was superseded by five number of folds.

The most optimum features discussed in the report were selected from each of the experiments performed to ensure maximum possible accuracy of classification. After the set of features were finalised they were now used for dialect classification on the available datasets. The following observations were made about the performance of the selected feature sets on one text dependent and another text independent data set:

**Table 8.** Various datasets and their highest accuracy of classification

| Dataset | Feature Name | No. of features | Accuracy (%) |
|---|---|---|---|
| Text Dependent | MFCC+Chroma+Others | 114 | 96.39 |
| Text Independent | MFCC+Chroma+Others | 114 | 87.115 |

The results during the initial stages of the experiment where over-fitting occurred with folds=10.

The first comparison was made to select a set of derivatives of chroma features. This includes CENS (chroma energy normalised statistics) and CRP (chroma DCT-reduced log pitch). The result of the experiment were as follows, accuracy is mentioned in (%):

**Table 9.** Comaprison between various derivatives of chroma features for dataset 1(Text dependent) and dataset 2(Text independent)

| Feature Name | Dataset 1-Accuracy(%) | Dataset 2-Acuuracy(%) |
|---|---|---|
| CENS | 68.333 | 61.718 |
| CRP | 69.0 | 60.93 |

**Table 10.** Comparison between various features for dataset-1(text dependent), dataset-2(text indeoendent) and TIMIT dataset

| Feature Name | Dataset 1-Accuracy(%) | Dataset 2-Acuuracy(%) | TIMIT-Accuracy(%) |
|---|---|---|---|
| MFCC(13) | 96.562 | 98.33 | 97.2 |
| MFCC(rasta) | 92.83 | 87.187 | 93.915 |
| FDLP-s | 90.0 | 85.781 | 21.0 |
| Chroma(original) | 96.167 | 94.062 | 97.2 |
| Chroma(updated) | 95.667 | 91.876 | 97.6 |
| MFCC+Chroma(updated)+Others | 96.667 | 93.281 | 95.9 |
| FDLP-s+MFCC+Others | 96.718 | 97.187 | 20.487 |

## V. CONCLUSION

### A. Summary

Two different sets of features were constructed which were capable to classify a given speech signal into the dialect class it belongs to for American English language. Both these feature sets are useful on different types of datasets. In real life situations it is very unlikely to encounter such a dataset for training and testing, thus limiting its usage. Although it can be used in PDAs (Personal Digital Assistant) where the commands are limited and hence can be considered to be text dependent.

In the experiments it was found that FDLP-s are a great alternative of MFCC features. Using these set of features dialect class of any given speech sample can be found with high accuracy.

### B. Limitations

The experiment had various limitations and attempts are needed in the process of overcoming them. The features selected were language specific and won't work for any given language with good classification accuracy. The third dataset having very similar dialect classes was not classified efficiently by the selected feature sets.

### C. Future Scope

Some of the future works include finding a feature set that are language independent and hence are able to classify the dialects of any given language. It would be quite difficult to achieve as there are various different kinds of languages and a general dialect classifier should first identify the language (or at least the type of language) before attempting to properly classify it. Also the set of features finalized should be fixed for both text dependent and text independent datasets and should give a fairly good accuracy for all the cases.

## VI. REFERENCES

1. Giannakopoulos T, ”A method for silence removal and segmentation of speech signals, implemented in MATLAB”.
2. Muller M, and Sebastian E, ”CHROMA TOOLBOX: MATLAB IMPLEMENTATIONS FOR EXTRACTING VARIANTS OF CHROMA-BASED AUDIO FEATURES”, 2011 .
3. David Gerhard, ”Audio Signal Classification: History and Current Techniques”, Technical Report TRCS 2003-07 November, 2003, p. 22.
4. Sriram Ganapathy, Samuel Thomas, and Hynek Hermansky, ”Static and Dynamic modulation spectrum for speech recognition”, ISCA, (2009)2823-2826.
5. Parwinder Pal Singh, and Pushpa Rani, ”An Aproach to extract features using MFCC”, IOSCR Journal of Engineering(IOSRJEN), Vol.04, August. 2014, Issue 8.
6. Kamil Aida-zade, Anar Xocayev, and Samir Rustamov, ”Speech recognition using Support Vector Machines”, Application of Information and Communication Technologies (AICT), 2016 IEEE 10th International Conference, 27 JULY 2017.
7. Manikandan J, and Venkataramani B, ”Design of a real time automatic speech recognition system

using Modified One Against All SVM classifier", ELSEVIER, 24 June 2011.

8. Yun Le., and John H. L. Hansen., "Dialect Classification via Text-Independent Training and Testing for Arabic, Spanish, and Chinese", IEEE Transactions On Audio, Speech, and Language Processing, Vol.19, No. 1, January (2011). 8
9. Esra J. Harfash, Abdul-kareem A. Hassan, "Automatic Arabic Dialect Classification", International Journal of Computer Applications (0975 8887), Volume 176 No.3, October 2017.
10. Mathworks, Inc.Matlab, 2014.