Assignment 4 Report: Dialect Classification

Ajay Ramesh
IMT2017502

Pratyush Nandi
IMT2017518

I. Overview

A dialect is a variety of a spoken language having specific linguistic features of vocabulary, grammar, and pronunciation that distinguish it from other varieties of the same language. Dialects commonly represent the changing speaking patterns observed in the group of native speakers belonging to some particular region. These dialects are responsible for failure of Automatic Sound recognition systems. So Dialect Classification is an interesting Signal-Processing-Pattern-Recognition problem. This report summarizes the extraction of features and classification of nine British English dialects.

II. Motivation

The first step is to extract features that can be used to identify linguistic content and discard other material like background noise etc. The sounds generated by humans are filtered by the shape of our vocal tract, including the tongue, teeth etc. The shape of these organs determine what sound is produced. Determining this shape would give us a good representation of the actual phoneme being produced. A phoneme is a unit of sound that distinguishes one word from another in a particular language. It is the smallest unit in a word. The shape of the vocal tract manifests itself in the power spectrum of the sound it produces. Thus, analysing the power spectrum would give us a good idea of the phoneme produced.

III. Data Exploration

The speech corpus IViE (Intonational Variation in English) dataset was used. The dataset contains nine dialects of British English, spoken across the nine various regions of the British Isles. The speakers include both male and female. The speakers are speaking sentences and thus, it includes various intonations and style variations. The audio was sampled at a rate of 16kHz. A audio recording of dialect 1 is shown in Fig 1. In order to capture the unique intonations of a dialect, the audio was divided into frames, and features were extracted for each frame.

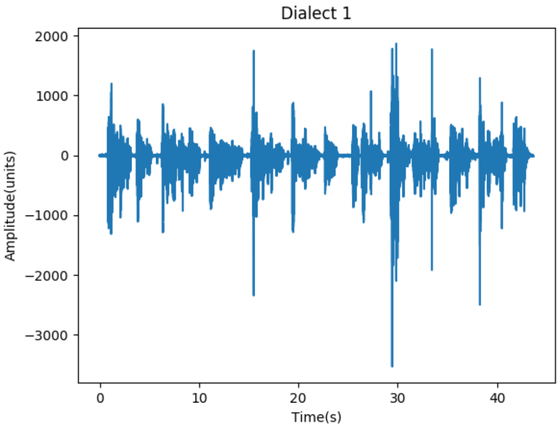

Fig. 1. Audio recording of dialect 1

IV. Feature extraction

A. Mel Frequency Cepstral Coefficents

MFCCs are features commonly used in Automatic Speech Recognition. The are computed as described -

1) The audio signal is framed. This is done to ensure that the signal does not change much during this short period. If the frame is too short, sufficient samples will not be available and if the frame is too long, the signal may change a lot during this period. Thus, an optimal frame length is required for best results. We have used a frame length of 25ms. In addition, to avoid sharp transitions, the hamming window was used with 10ms overlap between successive windows.
2) The power spectrum of each frame is computed. This is motivated by the human cochlea which vibrates depending on frequency of incoming sounds and fires corresponding nerve fibres.
3) The human ear cannot differentiate between very closely spaced frequencies and this becomes harder as the frequency increases. Thus, the energy of a clump of frequencies is considered, to give an idea of the energy of various frequency regions. The width of the clumps can be obtained by the Mel scale. The Mel scale relates actual frequency to the perceived frequency. The formula to convert to Mel scale is - $M(f) = 1125ln(1+f/700)$. These clumps of filters are called Mel filterbanks. We have used 13 filterbanks resulting in 13 features.
4) The logarithm of the filterbank energies is computed. This is because the human ear does not percieve loudness on a linear scale, but a log scale.
5) The DCT (Dicrete Cosine Transform) of the log of filterbank energies is computed to give the Mel filter coefficient features.

B. Features at the audio level

The MFCC features are computed for each frame of an audio. The structure of the features is shown, where an audio consists of N frames with 13 features each.

frame	feature1	feature2	.	.	feature13
1	.	.	.	.	.
2	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	.	.	.	.	.

In order to aggregate features at the audio level, feature vector for an audio file was computed by the mean of the feature vectors of the N frames constituting the audio. The python_speech_features tool box was used to extract the MFCC features. This constituted the data which was used to train the classifiers. The MFCC coefficients which form the feature vector for two different dialects are shown in Fig 2.

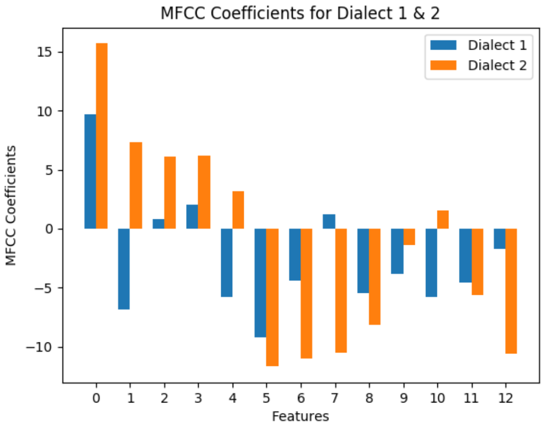

Fig. 2. Feature vector for two dialects

V. Classification

We used three classifiers - K-Nearest Neighbors, Logistic Regression and SVM with linear kernel. The dataset was split as 70% training and 30% testing. The training data was normalised to ensure accurate classification, since algorithms such as the K-nearest-neighbors use euclidean distance for distance computation.

VI. Results

The average accuracy was computed for 50 iterations while shuffling the dataset. The deviation of accuracy over the iterations was low, indicating that the model has not been overfitted. The KNN Classifier produced the best results with an average accuracy of 96.81% with a standard deviation of 2.7%. The following table summarises the results:

KNN(k=3)	KNN(k=5)	SVM	Logistic Regression
96.81%	95.22%	86.8%	70.22%

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. The confusion matrix of for the KNN classifier is shown in Fig 3. The confusion matrix tells us about the recall, precision and accuracy of the model. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

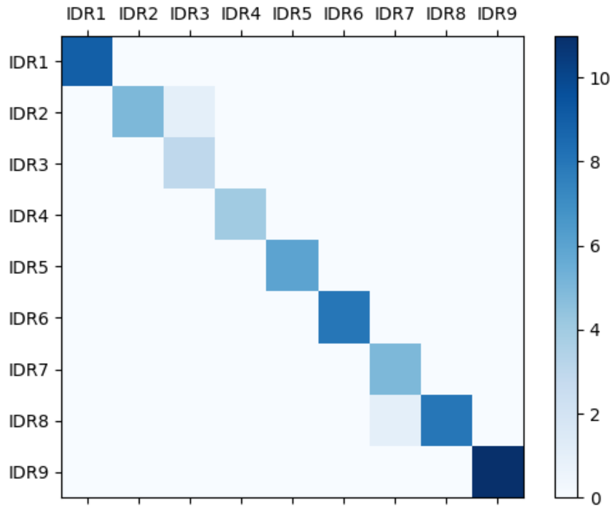

Fig. 3. Confusion matrix for the KNN classifier

VII. Conclusion

The Mel Frequency Cepstral Coefficients are a very useful set of features for automatic speech recognition. The why and how regarding these coefficients is indeed fascinating. They have been successfully used to classify various English dialects with a good accuracy. It can also be used to classify music, such as in terms of different genres etc.

VIII. References

1) Nagaratna B. Chittaragi, Shashidhar G. Koolagudi, ”Acoustic features based word level dialect classification using SVM and ensemble methods”. https://ieeexplore.ieee.org/document/8284315/authorsauthors
2) Speech Recognition — Feature Extraction MFCC PLP. https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9
3) Mel Frequency Cepstral Coefficient (MFCC). http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/