Dialect Classification

Nithin Raj
nithin.raj@iiitb.org
IMT2017511

Ronak Doshi
ronakvipul.doshi@iiitb.org
IMT2017523

***Abstract*—In this paper, a dialect classification system is proposed by using acoustic characteristics of speech signals. Dialects mainly represent the different pronunciation patterns of any language. Dialectal cues can exist at various levels such as phoneme, syllable, word, sentence and phrase in an utterance. Phoneme level dialectal traits are extracted to recognize dialects since every phoneme exhibits significant dialect discriminating cues. Intonational Variations in English (IViE) speech corpus recorded in British English has been considered. The corpus includes nine dialects which cover nine distinct regions of British Isles. Acoustic properties such as spectral features are derived from the passage audio files to construct the feature vector. Further, four different classification algorithms such as logisitc regression, support vector machine (SVM), k-nearest neighbors and random forest classification are used to extract the prominent patterns that are used to discriminate the dialects.**

***Index Terms*—Spectral Features, Phoneme, Logistic Regression, SVM, KNN, Random Forest**

I. Introduction

In this paper, the dialect classification system is proposed for nine dialects of British English. Since it is observed that acoustic characteristics for dialect discriminations are rich enough for the phoneme-level divisions of words, the proposed system is highly focused on processing and extracting phonemes. Spectral features such as cepstral coefficients (MFCC), delta coefficients, delta-delta coefficients, and spectral flux have been computed based on their efficiency in discriminating dialects. Further, to demonstrate the dialect classification performances, four distinct classification methods namely, individual classifiers such as logistic regression, SVM, KNN and Random Forest techniques have been utilized and the performances are compared. The system performance is also evaluated for the possible combinations of features.

II. Data Exploration

The IViE corpus contains recordings of nine urban dialects of English spoken in the British Isles. Recordings of male and female speakers were made in **London, Cambridge, Cardiff, Liverpool, Bradford, Leeds, Newcastle, Belfast** in Northern Ireland and Dublin in the Republic of Ireland. Three of the speaker groups are from ethnic minorities: we have recorded bilingual Punjabi/English speakers, bilingual Welsh/English speakers and speakers of Carribean descent. These **9** dialects have **67** audio files each. The audio samples are narrations of a Cinderella passage. Each audio file is roughly 1 minute long, sampled at 22 kHz. This gives us enough data to extract spectral information which will help us in the classification.

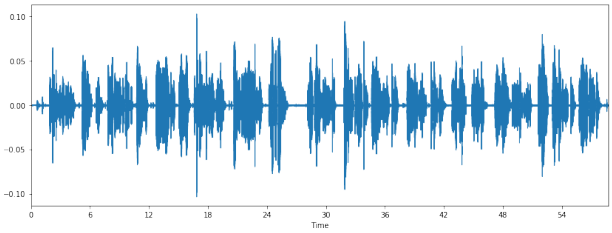

Fig. 1. Waveform for a sample passage of the Belfast dialect over time.

Because phonemes are part of words, we need to break down our audio samples into small parts for analysis. For this, we break down the audio signal into small parts which approximately contain enough spectral content on how different phonemes make up each word. Each small part would be one frame. Statistically, the length of one frame is usually around **25ms**. Anything more or less than this makes the spectral content of the frame ambiguous. While breaking down our audio sample into frames, we make sure there is some overlap present between them to find a correlation between adjacent phonemes. This overlap is called stride. Again, statistically, this stride is taken to be around **10ms**. Anything less won't be useful and anything more would increase overfitting. This is done because, in dialects, the temporal positioning of phonemes in speech also plays a huge role. A phoneme may change to some other phoneme when placed adjacent to some other phoneme.

III. Feature Extraction

A. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are a decades old tool for representing human speech as it is perceived. The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. Broadly, there are 5 steps in MFCC calculation :

- Frame the signal into short frames.
- For each frame calculate the periodogram estimate of the power spectrum.

- Apply the mel filterbank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filterbank energies.
- Take the DCT of the log filterbank energies.

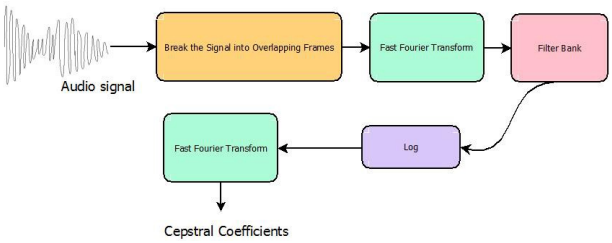

Fig. 2. Schematic representation of the steps necessary to create MFCCs from an audio signal.

1) Frame Level Breakdown: An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much, statistically. This is why we frame the signal into 25ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame. Also, there is an overlap of 10ms between each frame to capture phoneme transition.

2) Power Spectrum Calculation: This is motivated by the human cochlea which vibrates at different spots depending on the frequency of the incoming sounds. Depending on the location in the cochlea that vibrates (which wobbles small hairs), different nerves fire informing the brain that certain frequencies are present. Our periodogram estimate performs a similar job for us, identifying which frequencies are present in the frame.

3) Applying Mel Filterbank: In particular the cochlea can not discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. For this reason we take clumps of periodogram bins and sum them up to get an idea of how much energy exists in various frequency regions. This is performed by our Mel filterbank: the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. As the frequencies get higher our filters get wider as we become less concerned about variations. We are only interested in roughly how much energy occurs at each spot. The Mel scale tells us exactly how to space our filterbanks and how wide to make them.

4) Logarithm of Filterbank Energies: Once we have the filterbank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the percieved volume of a sound we need to put 8 times as much energy into it. This means that large variations in energy may not sound all that different if the sound is loud to begin with. This compression operation makes our features match more closely what humans actually hear. The logarithm allows us to use cepstral mean subtraction, which is a channel normalisation technique.

5) DCT of Log Filterbank Energies: The final step is to compute the DCT of the log filterbank energies. There are 2 main reasons this is performed. Because our filterbanks are all overlapping, the filterbank energies are quite correlated with each other. The DCT decorrelates the energies which means diagonal covariance matrices can be used to model the features. 20 DCT coefficients are computed, discarding the rest.

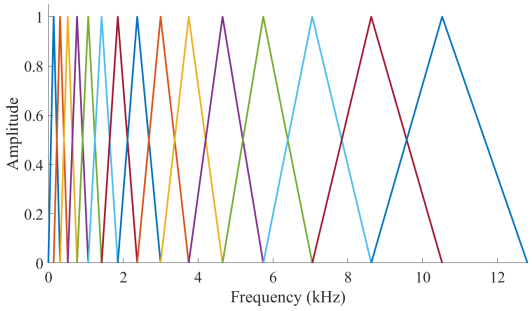

Fig. 3. Mel scale filterbanks applied to power spectrum (13 filterbanks out of 20 used shown)

Thus, the output of the MFCCs computation results in a 20-dimensional feature vector for each frame of the audio file. The mean of all the frames is computed and the resulting **20-dimensional feature** vector is the MFCC representative of that audio file.

B. Delta and Delta-Delta Coefficients

Till now we've worked on individual phoneme spectral content but dialects also have different velocities and acceleration of transition between phonemes. We noticed that the speech is faster in IDR3 compared to speech in IDR2. We can create delta coefficients (velocity) and delta-delta coefficients (acceleration) from the MFCC feature vector itself to learn these features. The mean is computed the way described in MFCC feature vector computation. This gives us **40 features** (20 delta features and 20 delta-delta features) for each audio file.

C. Spectral Flux

The spectral change between two consecutive frames is measured by spectral flux feature. It is possible to decide between two sounds whether two sounds are similar or not through feature called as a timbre. Spectral flux measures the quick changes occurring in the power spectrum of a signal. The mean value of spectral flux is computed by taking the average value of spectral flux for all frames of that audio file. This gives us **1 feature**.

Thus, for the classification, we have extracted **61 features** from the dataset.

IV. Model

After extracting all the necessaries features, we tried multiple models to get an idea of which performs better. In total we tested the data-set across 6 different models:

1) Multinomial Logistic Regression
2) Support Vector Machine (SVM)

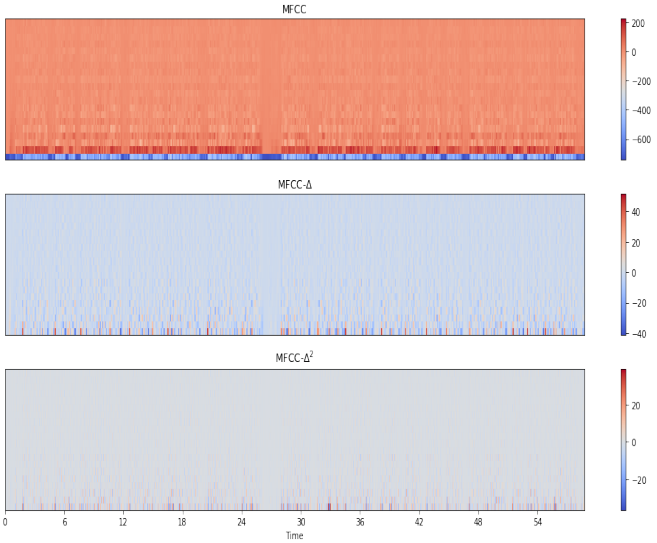

Fig. 4. Visual representation of the 20 MFCCs, 20 Delta, and 20 Delta-Delta coefficients for a sample passage of the Belfast dialect.

3) K-Nearest Neighbors (KNN)
4) Random Forest

All these models were trained on three different dataframes. Each data-frame contained specific set of features that were extracted earlier.
Features in each of the data-frame were:

- **Dataframe 1:** 20 MFCC features
- **Dataframe 2:** 20 MFCC features, 20 Delta features, 20 Delta-Delta features and 1 spectral flux feature
- **Dataframe 3:** 20 MFCC features and 1 spectral flux feature

TABLE I
ACCURACY OF THE MODEL CORRESPONDING TO THAT DATAFRAME

	Dataframe 1	**Dataframe 2**	**Dataframe 3**
Logistic Regression	0.845	0.797	0.854
SVM	0.98	0.825	0.988
KNN with k=3	0.976	0.829	0.983
KNN with k=5	0.953	0.816	0.958
Random Forrest	0.965	0.94	0.975

One important thing to notice is that accuracy of all the models decrease drastically when trained on dataframe 2 than on dataframe 1. From this we can conclude that adding more features is decreasing the accuracy. Instead of helping the model to predict with better accuracy is in-fact making the model more confused. Thus, we can say that velocity and acceleration features (delta and delta-delta coefficients) don't show much variation across all dialects, thus affecting the accuracy of our predictions.

On the other hand, we can see that the accuracy is increased when all the models are trained on dataframe 3 than on dataframe 1. This shows that spectral flux is a very vital feature showing variance across dialects, thus helping us improve the accuracy of our predictions.

V. CONCLUSION

After analyzing all the models, the model with highest accuracy came out to be Support Vector Machine (SVM) when feeded with features from MFCC and Spectral Flux. SVM also showed very low standard deviation. This means that the model isn't overfitting the training data. The SVM model also showed very high precision, recall, and f1-score for each of the features.

Accuracy of the model is 98.8% with standard deviation in accuracy of 0.0129.

REFERENCES

[1] "Librosa feature extraction docs," https://librosa.github.io/librosa/feature.html.
[2] "Speech processing for machine learning-haytham fayek," https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.
[3] "Dummy's guide to mfcc-medium article," https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd.
[4] "Mfcc tutorial," http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.
[5] N. B. Chittaragi and S. G. Koolagudi, "Acoustic features based word level dialect classification using svm and ensemble methods," in *Accepted in 2017 Conference on Contemporary Computing*, August 2017.