Acoustic Features Approach for Arabic Dialects Recognition

Ziad Adeela, Maher Khdeir

*#Electrical and Computer Engineering Dept. Birzeit University*zeiad.suliman@gmail.com

Maher.Khdeir@gmail.com

16 Dec 2017

***Abstract*— This report is an approach for recognition of arabic dialects (Egypt, Gulf, Levantine, Standard Arabic, and Magharibi) using acoustic features i.e. MFCC features using GMM for classification.**

# Introduction

1. **Mel-frequency Cepstral Coefficients.**

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.[1]

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the sound. They are commonly derived as follows:

* Take the Fourier transform of (a windowed excerpt of) a signal.
* Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
* Take the logs of the powers at each of the mel frequencies.
* Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
* The MFCCs are the amplitudes of the resulting spectrum.ral representation of the audio clip .

1. **Gaussian Mixture Model**

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the means of the latent Gaussians.[2]

The BIC criterion can be used to select the number of components in a Gaussian Mixture in an efficient way. In theory, it recovers the true number of components only in the asymptotic regime (i.e. if much data is available and assuming that the data was actually generated from a mixture of Gaussian distribution). Note that using a Variational Bayesian Gaussian mixture avoids the specification of the number of components for a Gaussian mixture model.

The main difficulty in learning Gaussian mixture models from unlabeled data is that it is one usually doesn’t know which points came from which latent component (if one has access to this information it gets very easy to fit a separate Gaussian distribution to each set of points). Expectation-maximization is a well-founded statistical algorithm to get around this problem by an iterative process. First one assumes random components (randomly centered on data points, learned from k-means, or even just normally distributed around the origin) and computes for each point a probability of being generated by each component of the model. Then, one tweaks the parameters to maximize the likelihood of the data given those assignments. Repeating this process is guaranteed to always converge to a local optimum.

# II. PROCEDURE

1. **Training Process.**

First step of this system, is to calculate the feature vector that will be used in the classification Process. The MFCCs were chosen as this feature vector. A 12 coefficients were taken, and then the energy was appended. Since these 13 features are not enough, the delta and double delta were added to these features, therefore the feature vector now has 39 dimensions.

After that, A GMM model will be built for every class in out dataset. And can be achieved by collecting all 39 features from every frame, for all wav files in that class together, then gmm training will be applied. To determine the best number of components for each class, a configurable range (a, b) can be specified, then after building models from “a” components to “b” components, the best one using BIC(Bayesian information criterion) criterion will be chosen. Note that the training process will train the gmms. The initial values for means, covariances and weights will be chosen using k-means algorithm, then the EM algorithm to specify the final values of them. The maximum number of iterations for EM were set to 100. And the covariance type {‘full’, ‘tied’, ‘diag’, ‘spherical’}that will be taken is the one that achieves the lowest BIC. After having the model for each class, they will be stored to avoid making the training process again.

1. **Testing Process.**

After creating the training models. Now is the time to evaluate the system. To do that, the mfcc will be generated for testing files. Then for each file, it will be applied for every gmm to calculate the weighted log probabilities for each sample. Then the sum of all samples will be done. After that, the with maximum sum of weighted log probabilities will be taken as the class for this file.

# III. Date set and Results

A date set of 400 wav file were taken for each class. 80% of them were taking for training, and 20% taken for testing. The dataset has 5 classes of arabic dialects which are (EGY, GLF, LAV, MSA, and NOR).

The best number of gaussians from range (50 to 79), and the best covariance type were chosen for every class:

|  |  |  |
| --- | --- | --- |
| Class | Number of Gaussians | Covariance type |
| EGY | 67 | Full |
| GLF | 63 | Full |
| LAV | 68 | Full |
| MSA | 66 | Full |
| NOR | 67 | Full |

The results are represented as the following confusion matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | EGY | GLF | LAV | MSA | NOR |
| EGY | 45 | 4 | 6 | 6 | 19 |
| GLF | 5 | 48 | 0 | 16 | 11 |
| LAV | 13 | 0 | 14 | 18 | 35 |
| MSA | 23 | 0 | 3 | 48 | 6 |
| NOR | 9 | 5 | 2 | 8 | 56 |

And the evaluation metrics are listed below:

Precision = 56.3%

Recall = 52.7%

F-measure = 51.3%

Accuracy = 55.2%

IV. Conclusion

In this report, an Arabic dialects recognition were done by using MFCC features, 12 first coefficient appended with energy, then the delta and double delta, which make them a 39 dimensional feature vector. And GMM chosen as a classifier. The GMM initial values were chosen using k-means, the the final values were chosen by EM algorithm with 100 as the max number of iteration.

The system was applied on 2000 wav file, 400 for each class. 80% were taken for training, and 20% for testing. The results was as following, Precision = 56.3% ,Recall = 52.7%,F-measure = 51.3%, and Accuracy = 55.2%.

Therefore, the Acoustic Features are not the best features for recognition Arabic Dialects. Instead, ASR can be applied to speech, then processing and features could be taken from the text itself, like n-gram.

References

[1]<https://en.wikipedia.org/wiki/Mel-frequency_cepstrum> Accessed on 16/12/2017

[2] <http://scikit-learn.org/stable/modules/mixture.html#gaussian-mixture-models> Accessed on 16/12/2017