Bangladeshi Dialect Recognition using MFCC, Delta, Delta-delta and GMM

Submitted by:

Ruhul Amin	**ID#2011-2-60-027**
Zahida Rahman	**ID#2011-2-60-028**

Supervised by:

Dr. Shaikh Muhammad Allayear

Assistant Professor
Department of Computer Science and Engineering
East West University

A Project Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelors of Science in Computer Science and Engineering to the Department of Computer Science and Engineering
At the

East West University
Dhaka, Bangladesh
September, 2015

Abstract

Automatic recognition systems are commonly used in speech processing to classify observed utterances by the speaker identity, dialect and language. A lot of research has been performed to detect speeches, dialects and languages of different region throughout the world. But the work on dialects of Bangladesh is infrequent to our research. These dialects, in turn, differ quite a bit from each other. In this paper, we propose a method to detect Bangladeshi different dialects which utilizes Mel-Frequency Cepstral Coefficients (MFCC), its Deltas and Delta-Deltas as main features and Gaussian Mixture Models (GMM) to classify characteristics of a specific dialect. Particularly we extract the MFCC, Deltas and Delta-Deltas features from the speech signal. Then they are merged together to form a feature vector for a specific dialect. GMM is trained using the iterative Expectation Maximization (EM) algorithm where feature vectors are served as input. This scheme is tested on 5 databases of 30 speech samples each. Speech samples are contained dialects of Borishal, Noakhali, Sylhet, Chittagong and Chapai Nobabgonj regions of Bangladesh. Experiments show that GMM adaptation gives comparable good performance.

Declaration

We hereby declare that, this project was done under CSE497 and has not been submitted elsewhere for requirement of any degree or diploma or for any purpose except for publication.

Ruhul Amin
ID#2011-2-60-027
Department of
Computer Science and Engineering
East West University

Zahida Rahman
ID#2011-2-60-028
Department of
Computer Science and Engineering
East West University

Letter of Acceptance

I hereby declare that this thesis is from the student's own work and best effort of mine, and all other source of information used have been acknowledge. This thesis has been submitted with my approval.

Dr. Shaikh Muhammad Allayear **Supervisor**
Assistant Professor
Department of Computer Science and Engineering
East West University

Dr. Shamim Hasnat Ripon **Chairperson**
Associate Professor and Chairperson
Department of Computer Science and Engineering
East West University

Acknowledgement

First of all we would like to thank Almighty Allah for giving us the strength and patience.

Our sincere gratitude goes to our parents for their all kindness and encouragement during our studies, without them this thesis would have been very hard to finish.

We would like to thank our supervisor Dr. Shaikh Muhammad Allayear for his utmost direction, constant helpful support and feedbacks.

We are extremely grateful to the voice sample providers without their cooperation; it could not be possible to accomplish this thesis.

We also thank the researchers for their works that help us to learn and implement automatic speech recognition system.

Last but not the least, we express gratitude to our friends, seniors and juniors for providing effective suggestions and supporting us especially Pronaya Prosun Das.

Abbreviation and Acronyms

GMM- Gaussian Mixture Model

MFCC- Mel-Frequency Cepstral Coefficients

EM- Expectation Minimization

FFT- Fast Fourier Transform

DFT- Discrete Fourier Transform

PPRLM - Parallel Phone Recognition Language Modeling

SVM- Support Vector Machine

GLDS - Generalized Linear Discriminant Sequence

HMM- Hidden Markov Model

LID - Language Identification

SDC - Shifted Delta Costar

UBM- Universal Background Model

DCT- Discrete Cosine Transform

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

Language is the best ways to communicate among the human being. However, variation in dialect causes misunderstanding in communication. Therefore detection of regional language is essential.

1.1 Problem Specification

Language detection is one of the foremost issues in the field of computer science. It is a solved problem worldwide. Many developed country have worked with language detection and succeeded with many software. But no initiative has been taken for regional language detection of Bengali language till now. We have a prosperous language Bengali. About 250 million native people speak Bengali except different people speak in different regional language in Bangladesh. For instance, Borishali (Barisal region), Noakhali (Noakhali region), Rongpore (Rangpur Region), Khulna (Khulna region), Mymonshingh (Mymensingh region), Sylheti (Sylhet region) and Chittagonian (Chittagong region) are major spoken dialects in Bangladesh. Although these languages are mutually intelligible with neighboring dialects of Bengali, they have lack of mutual intelligibility with the Bengali language and sometimes would not be understood by a native speaker of Standard Bengali. Hence, some of these dialects are sometimes considered as languages in their own right.

1.2 Motivation

Basically Bengali language varies due to some cultural impact. Sometime it varies a lot that it becomes quite complicated to understand which region a person belongs to and it is also very difficult for a person to identify all the regional languages. That is where we got our motivation to develop such a method which will detect different dialects. It is a part of artificial intelligence and considered as machine learning. Moreover it is a critical issue to train a system and teach how to detect the specific dialect accurately.

1.3 Proposed Method of this thesis

Our system uses the Gaussian Mixture Modeling (GMM) approach and it operates in two phrases. First one is training phrase, where it takes the speech utterances for a single dialect and converts them into feature vectors. A GMM is trained on the feature vectors to each dialect. Second phrase is recognition where an unknown utterance is compared to each of the GMMs. The likelihood that the unknown utterance was spoken in the same dialect as the speech used to train each model is computed, and the most likely model is determined as the hypothesized dialect [1]. Feature vectors consist of Mel Frequency Cepstral Coefficients (MFCC), its Deltas and Delta-Deltas. To achieve the above mentioned objectives the methodology adopted is explained by the following flowchart shown in Figure 1.

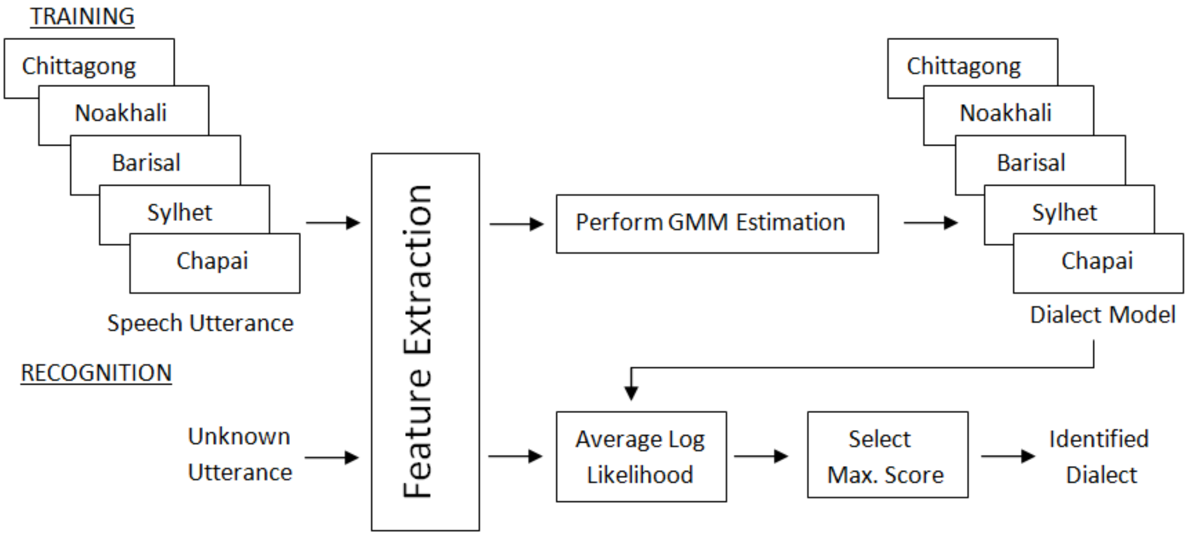


Figure 1: Block diagram of Dialect Recognition System

The input audio clip is first pre-processed. In pre-processing the raw sample file with duration of 90secs is clipped. After this, the clipped audio is converted into .wav format. Then, MFCCs, Deltas and Delta-deltas [2-3] were extracted from the formatted audio for all dialects. Then classification of the language is done using GMM.

The database of 30 samples is created for Borishali, Noakhali, Sylheti, Chittagonian and Chapai Nobabgonj dialect each due to training and recognition phrase. In each database, 15 audio clips have male voice and 15 clips are recorded with female voice. 8 samples of each language are used for training and rest 22 is used for testing.

1.4 Structure of this thesis

Step by step approaches to Dialect Recognition is mentioned in the following paragraph. In Section 2 other approaches of language detection are being described elaborately, Section 3 describes how MFCC is used for the feature extraction. Next Section 4 describe in detail approaches of GMM properly. The results of the system used for Dialect Recognition are presented in Section 5. Additionally here we will have a discussion on the sample collection. Conclusions and future work can be found in Section 6.

Chapter 2

Related Works

A good deal of effort has been made in the recent past by researchers in their attempt to come up with computational intelligence models with an acceptable level of classification accuracy.

2.1 Automatic language identification with discriminative language characterization based on svm

[4] presented three mainstream approaches including Parallel Phone Recognition Language Modeling (PPRLM), Support Vector Machine (SVM) and the general Gaussian Mixture Models (GMMs). The experimental results showed that the SVM framework achieved an equal error rate (EER) of 4.0%, outperforming the state-of-art systems by more than 30% relative error reduction. Also, the performances of their proposed PPRLM and GMMs algorithms achieved an EER of 5.1% and 5.0% respectively.

2.2 SVM-UBM based automatic language identification using a vowel-guided segmentation

Support Vector Machines (SVMs) were presented by [5] by introducing a sequence kernel used in language identification. Then a Gaussian Mixture Model was developed to do the sequence mapping task of a variable length sequence of vectors to a fixed dimensional space. Their results demonstrated that the new system yielded a performance superior to those of a GMM classifier and a Generalized Linear Discriminant Sequence (GLDS) Kernel.

2.3 Language identification using Gaussian mixture model tokenization

[6] presented a generalized technique by using GMM and obtained an error of 17%. In another related work, [10] described two GMM-based approaches to language identification that use Shifted Delta Costar (SDC) feature vectors to achieve LID performance comparable to that of the best phone-based systems. The approaches included both acoustic scoring and a GMM tokenization system that is based on a

variation of phonetic recognition and language modeling. The results showed significant improvement over the previously reported results.

2.4 Speaker verification using adapted gaussian mixture models

A description of the major elements of MIT Lincoln Laboratory's Gaussian Mixture Model (GMM)-based speaker verification system built around the likelihood ratio test for verification, using simple but effective GMMs for likelihood functions, a Universal Background Model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM were presented by [7]. The results showed that the GMM-UBM system has proven to be very effective for speaker recognition tasks.

2.5 A multiple-Gaussian classifier for language identification using acoustic information and PPRLM scores

[8] presented an additive and cumulative improvements over several innovative techniques that can be applied in a Parallel Phone Recognition followed by Language Modeling (PPRLM) system for language identification (LID), obtaining a 61.8% relative error reduction from the base system. They started from the application of a variable threshold in score computation with a 35% error reduction, then a random selection of sentences for the different sets and the use of silence models, then, compared the bias removal technique with up to 19% error reduction and a Gaussian classifier of up to 37% error reduction, then, included the acoustic score in the Gaussian classifier with 2% error reduction, increased the number of Gaussians to have a multiple-Gaussian classifier with 14% error reduction and finally, included additional acoustic HMMs of the same language with success gaining 18% relative improvement.

2.6 Dialect identification using gaussian mixture models

[9] evaluated the related problem of dialect identification using the GMMs with SDC features. Results showed that the use of the GMM techniques yields an average of 30% equal error rate for the dialects in one language used and about 13% equal error rate for the other one.

Chapter 3

Feature Extraction

The information in speech signal is actually represented by short term amplitude spectrum of the speech wave form. This allows us to extract features based on the short term amplitude spectrum from speech. The fundamental difficulty of speech recognition is that the speech signal is highly variable due to different speakers, speaking rates, contents and acoustic conditions. Mel Frequency Cepstral Coefficients (MFCC), Delta and Delta-Deltas have been used for our system.

The use of Mel Frequency Cepstral Coefficients can be considered as one of the standard method for feature extraction and they are widely used features for automatic speech recognition systems to transform the speech waveform into a sequence of discrete acoustic vectors [10-11]. The non-linear frequency scale used an approximation to the Mel-frequency scale which is approximately linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz. This is motivated by the fact that the human auditory system becomes less frequency-selective as frequency increases above 1 kHz [10]. In the sound processing, the Mel-frequency cepstrum 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-frequency scale [11]. Later we calculate the Deltas and Delta-Deltas which are actually the first and second derivative of MFCCs.

The procedure [2] is described as follows:

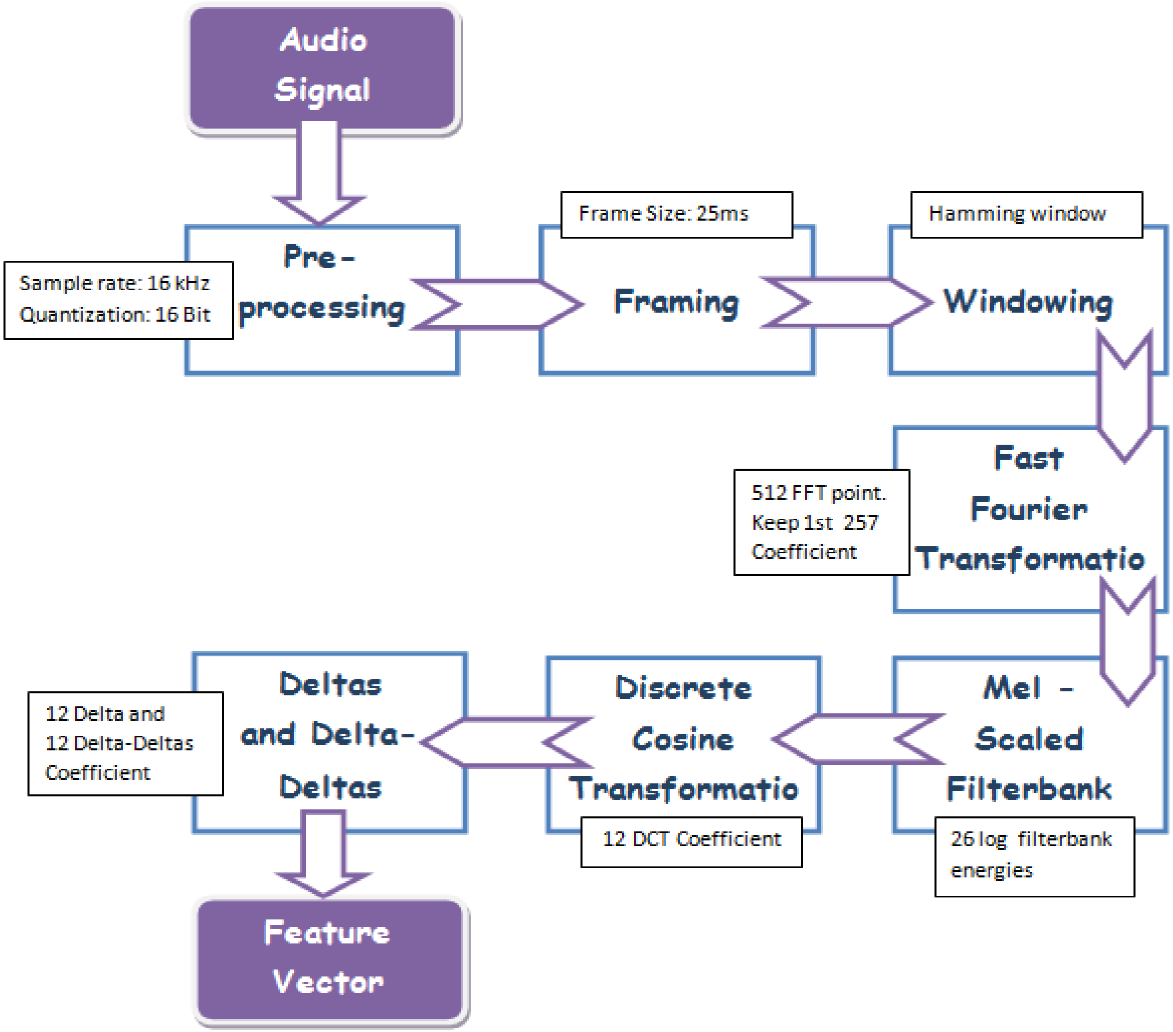


Figure 2: Block diagram of feature extraction: MFCC, Delta and Delta-Deltas

3.1 Pre-Processing

Speech signal is analog. In the first place analog electrical signals are converted to digital signals. This is done in two steps, sampling and quantization [11]. 16 kHz sample rate provides more accurate high frequency information. For most ASR applications, sampling rates higher than about 22 kHz is a waste. And 16Bit quantization is suitable for speech recognition.

3.2 Framing

An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much (when we say it doesn't change, we mean statistically i.e. statistically stationary, obviously the samples are constantly changing on even short time scales). That is why we frame the signal into 20-40ms 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.

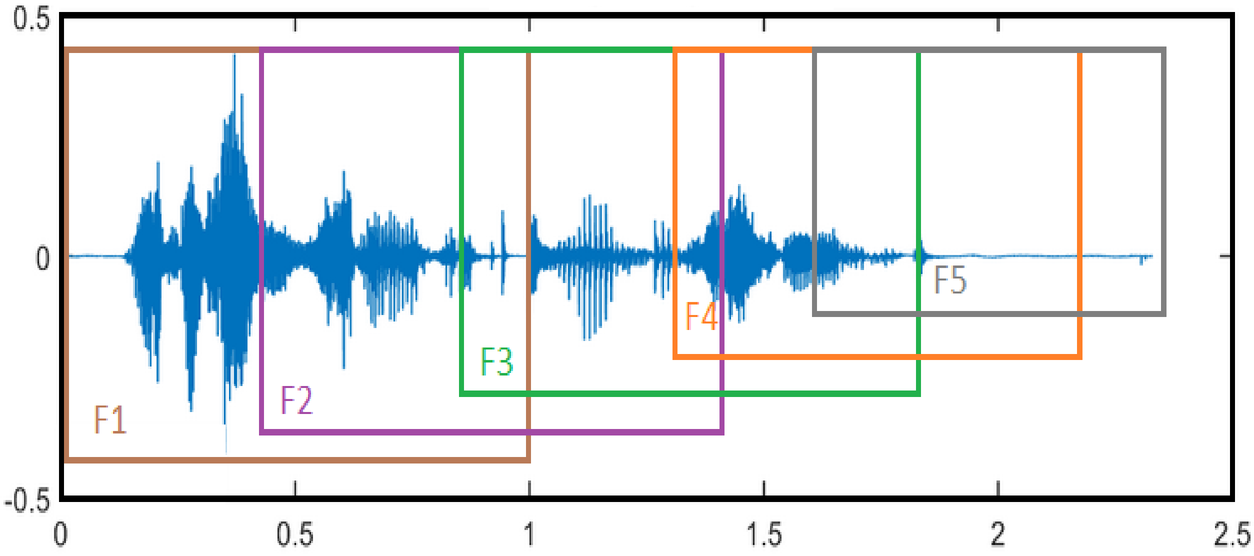


Figure 3: Framing

Frame the signal into 20-40 ms frames. 25ms is standard. This means the frame length for a 16 kHz signal is 0.025*16000 = 400 samples. Frame step is usually something like 10ms (160 samples), which allows some overlap to the frames. The first 400 sample frame starts at sample 0, the next 400 sample frame starts at sample 160 etc. until the end of the speech file is reached. If the speech file does not divide into an even number of frames, pad it with zeros so that it does.

3.3 Hamming Window

When we perform Fast Fourier Transformation (FFT) on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem and to reduce the signal’s discontinuity, the framed signal is passed through the Hamming window.

If the signal in a frame is denoted by $\boldsymbol{s(n), n = 0, \ldots N-1}$, N is size of the frame. Then the signal after Hamming windowing is

$$x(n) = s(n) * w(n,a) \qquad (1)$$

Where, $w(n)$ is the Hamming window defined by:

$$w(n,a) = (1-a) - a\cos\left(\frac{2\pi n}{N-1}\right); 0 \le n \le N-1 \qquad (2)$$

Different values of $\boldsymbol{a}$ corresponds to different curves

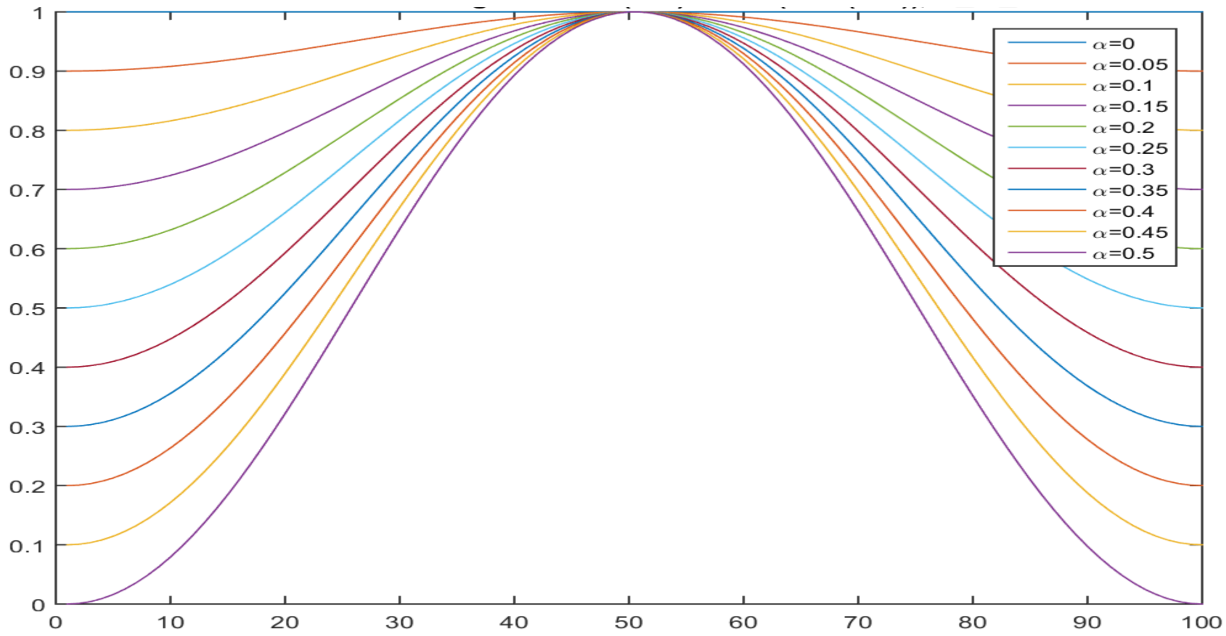
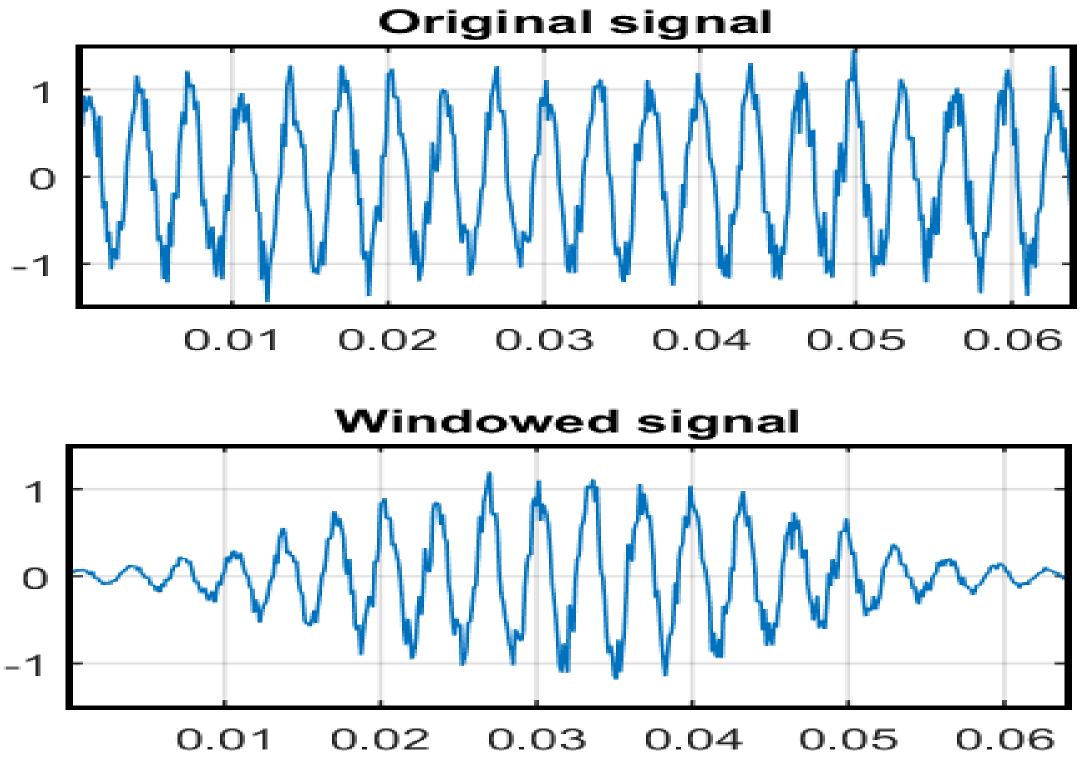


Figure 4: Generalized Hamming window of eq (2)

Hamming windows shown in figure 4.



.

Figure 5: Effect of windowing.

3.4 Discrete Fourier Transformation

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform DFT to obtain the magnitude frequency response of each windowed signal. The equation for DFT is:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}, \quad 0 \leq k \leq K-1 \qquad (3)$$

K is the length of the DFT. We would generally perform a 512 point FFT and keep only the first 257 coefficients.

The Fast Fourier Transformation is basically an efficient algorithm for computing the DFT. More specifically, FFT is the name for *any* efficient algorithm that can compute the DFT in about $O(n \log n)$ time, instead of $O(n^2)$ time.

The FFT convert each frame of N samples from time domain into frequency domain. Thus the components of the magnitude spectrum of the analyzed signal are calculated.

$$Y(\omega) = FFT[h(t) * x(t)] = H(\omega)X(\omega) \qquad (4)$$

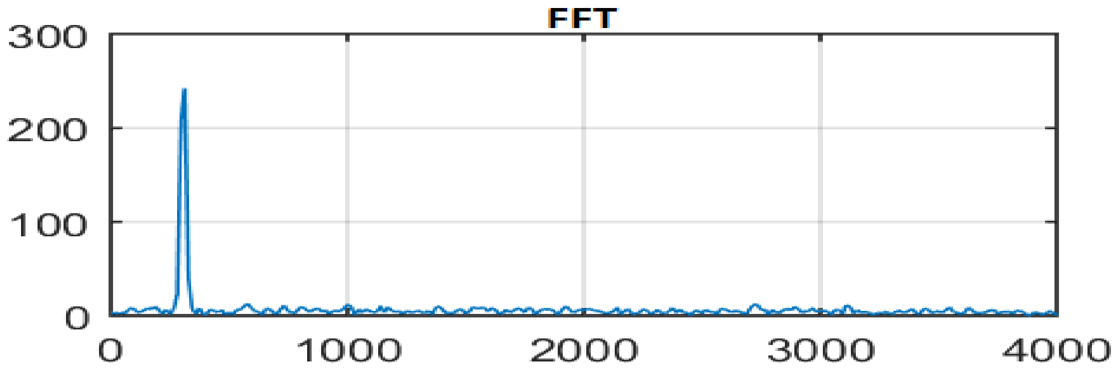


Figure 6: Power Spectrum

The periodogram-based power spectral estimate for the speech frame $X(n)$ is given by:

$$P(k) = \frac{1}{N}|X(k)|^2 \qquad (5)$$

We take the absolute value of the complex Fourier transform, and square the result. From here if power of each frame is bellow 1.2*mean power then these low power fremes are removed.

3.5 Mel-Scaled Filterbank

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The Mel-Scaled filterbank is then applied to each spectral block to convert the scale to a mel scale. The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \qquad (6)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right) \qquad (7)$$

The relationship between the mel and the linear frequencies is shown below:

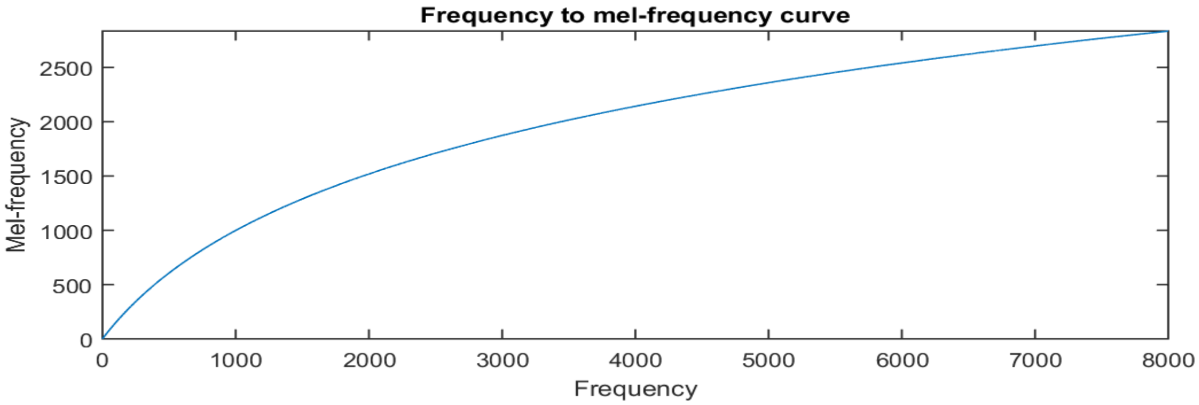


Figure 7: Frequency to mel-frequency curve

Set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components [12].

The following figure shows a traditional triangular mal-scaled filterbank:

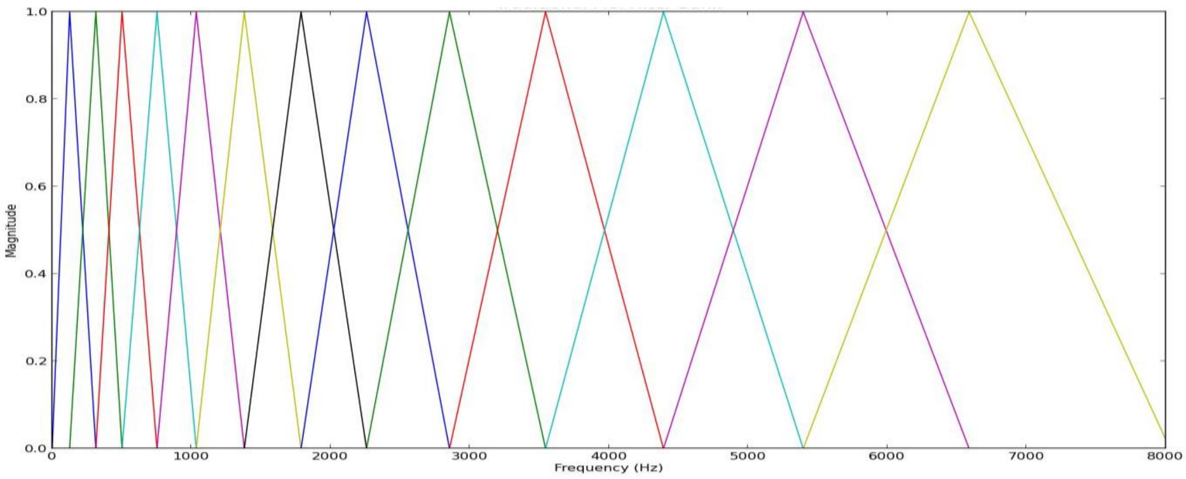


Figure 8: Traditional Mel Filter Bank

We apply a set of 20-40 (26 is standard) triangular filters to the periodogram power spectral estimate. Our filterbank comes in the form of 26 vectors of length 257. Each vector is mostly zeros, but is non-zero for a certain section of the spectrum. To calculate filterbank energies we multiply each filterbank with the power spectrum, then add up the coefficients. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filterbank.

The reasons for using triangular bandpass filters are two. First one is to smooth the magnitude spectrum such that the harmonics are flattened in order to obtain the envelop of the spectrum with harmonics. This indicates that the pitch of a speech signal is generally not presented in MFCC. As a result, a speech recognition system will behave more or less the same when the input utterances are of the same timbre but with different tones/pitch. and finally to reduce the size of the features involved.

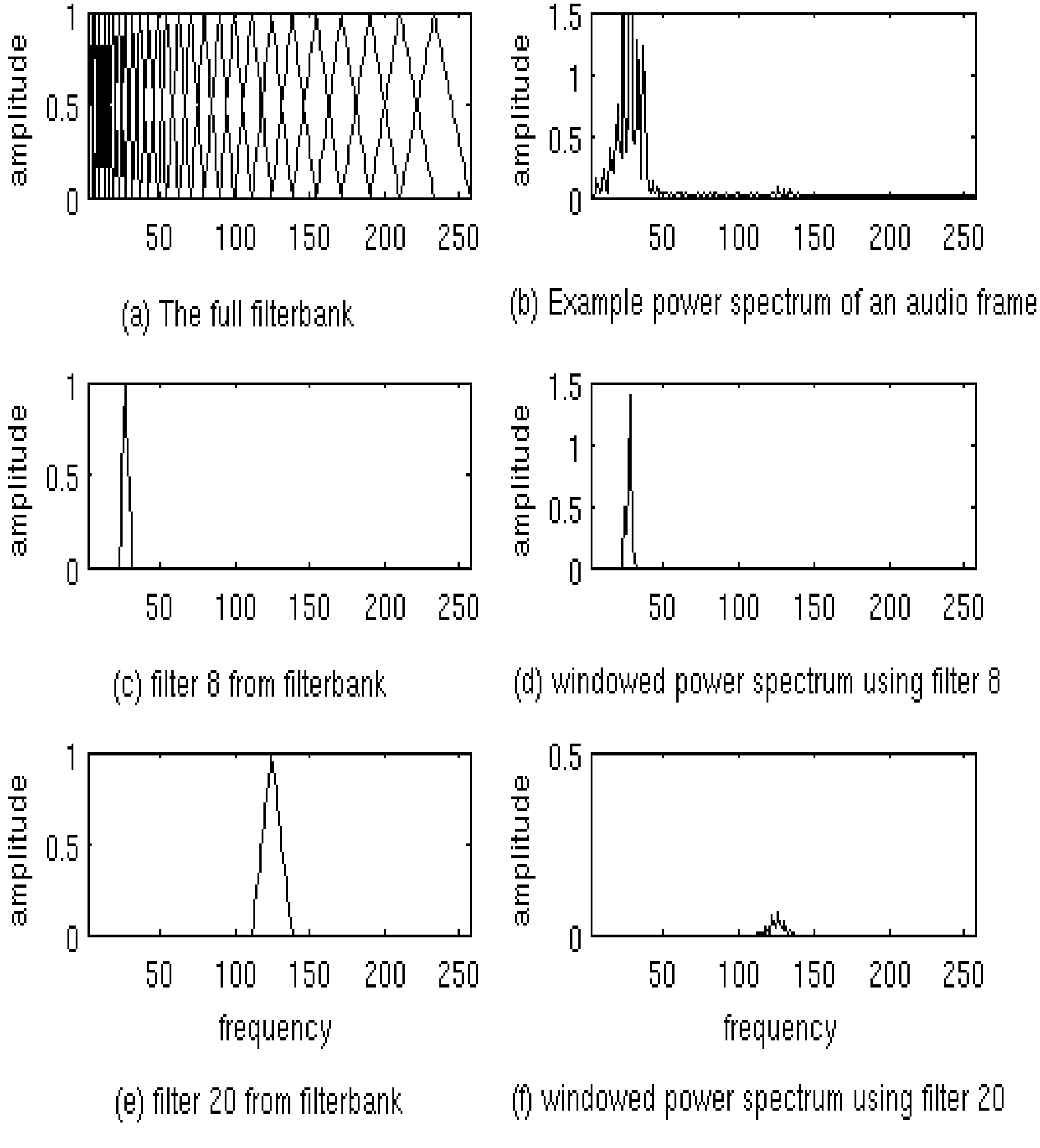


Figure 9: Triangular Bandpass filter over energy spectrum

3.6 Log Filterbank Energies

Once we have the filterbank energies, we take the logarithm of them. We take the *log* of each of the 26 energies. This leaves us with $26\,log$ filterbank energies. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the perceived 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 normalization technique.

3.7 Discrete cosine transform or DCT

In this step, we apply DCT on the 26 log energy E_k obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients (MFCC). The formula for DCT is shown next. DCT is chosen because it has excellent energy compaction. Eq (8) describing DCT is:

$$Y(k) = \sum_{l=0}^{L-1} \cos[\frac{\pi}{L}(l + \frac{1}{2})m]E_k; \;\; 0 \le m \le L-1 \qquad (8)$$

Where, L=26

But notice that only 12 of the 26 DCT coefficients are kept. This is because the higher DCT coefficients represent fast changes in the filterbank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them. MFCC alone can be used as the feature for speech recognition.

3.8 Deltas and Delta-Deltas

Deltas and Delta-Deltas also known as differential and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but it seems like speech would also have information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time. It turns out that calculating the MFCC trajectories and appending them to the original feature vector increases ASR performance by quite a bit (if we have 12 MFCC coefficients, we would also get 12 delta coefficients, and 12 delta-delta coefficients, which would combine to give a feature vector of length 36.

To calculate the delta coefficients, the following formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (9)$$

Where d_t is a delta coefficient, from frame *t* computed in terms of the static MFCCs c_{t+n} to c_{t-n}. A typical value for *N* is 2. Delta-Delta (Acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients.

Chapter 4

Models and Classifiers

There are lots of models and classifiers that have been used for speaker recognition and verification. Although, early classifiers for speaker recognition include non-parametric technique like VQ, DTW etc. now a day's classification methods for speaker recognition have centered on statistical approaches like HMM, GMM etc [13].

The structure and choice of a classifier depends on the application and the features used as well as the level of user cooperation, expected channels and recording devices, amount of speech data available for enrollment and detection and finally the requirement of recognition accuracy [13].

Over the last decade, the Gaussian Mixture model (GMM)[14]has become established as the standard classifier for text-independent speaker recognition. GMM often used to the speaker verification because this mode has good ability of recognition [15]. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped distributions [16]. GMMs have unique advantages compared to other modeling approaches because their training is relatively fast and the models can be scaled and updated to add new speakers with relative ease [17].

4.1 Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model[16].

A Gaussian Mixture Model for dialect recognition system served as the simplest algorithm for this study. As shown below, GMM dialect identification is motivated by the observation that different dialects have different sounds and sound frequencies. GMM-based classification has been applied at several sites [18-20].

The GMM approach to model the probability density function of a feature vector, $\bar{x}$, by the weighted combination of multi-variate Gaussian densities:

$$p(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \qquad (10)$$

With

$$b_i = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(\bar{x}-\overline{\mu}_i)^T \Sigma^{-1} (\bar{x}-\overline{\mu}_i)} \qquad (11)$$

Where λ is the model described by

$$\lambda = \{p_i, \overline{\mu}_i, \Sigma_i\} \qquad (12)$$

In eq. (10), *i* is the mixture index (1≤ *i*≤ M), p_i is the mixture weight such that

$$\sum_{i=1}^{M} p_i = 1 \qquad (13)$$

And $b_i(\bar{x})$ is a multi-variate Gaussian distribution defined by [21] the corresponding means $\overline{\mu}_i$ anddiagonal covariance matrices, $\sum_i$.

For each dialect, a GMM is created. At first MFCC, Delta and Delta-Delta features are extracted and combined together to form a feature vector of 36 dimensions from training speech spoken in the dialect *d* as described previously.

A scatter diagram of first two dimensions from a feature vector of a dialect is shown below:

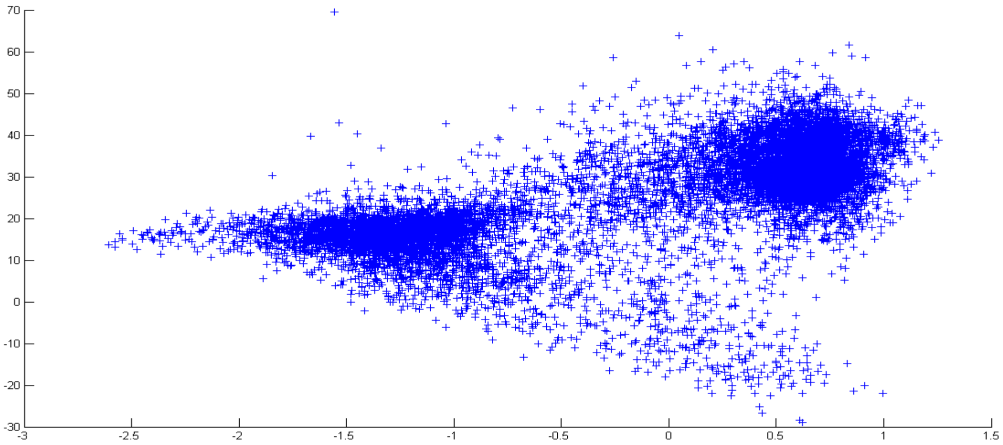


Figure 10: Scatter diagram of first two dimensions from a feature vector

We are using multiple speech samples from each dialect to train a GMM. So we get a set of feature vectors. Multiple iterations of the estimate-maximize algorithm are run by using the feature vectors as initial estimates for the means μ_i. For each stream, this process produces a more likely set of $\overline{\mu}_i, \sum_i$ and p_i [23-24].

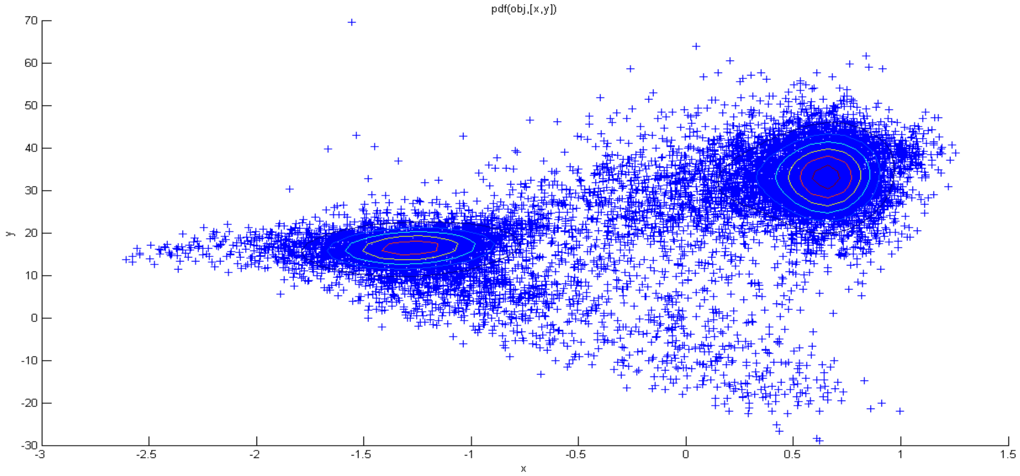


Figure 11: Scatter diagram of first two dimensions from a feature vector during training

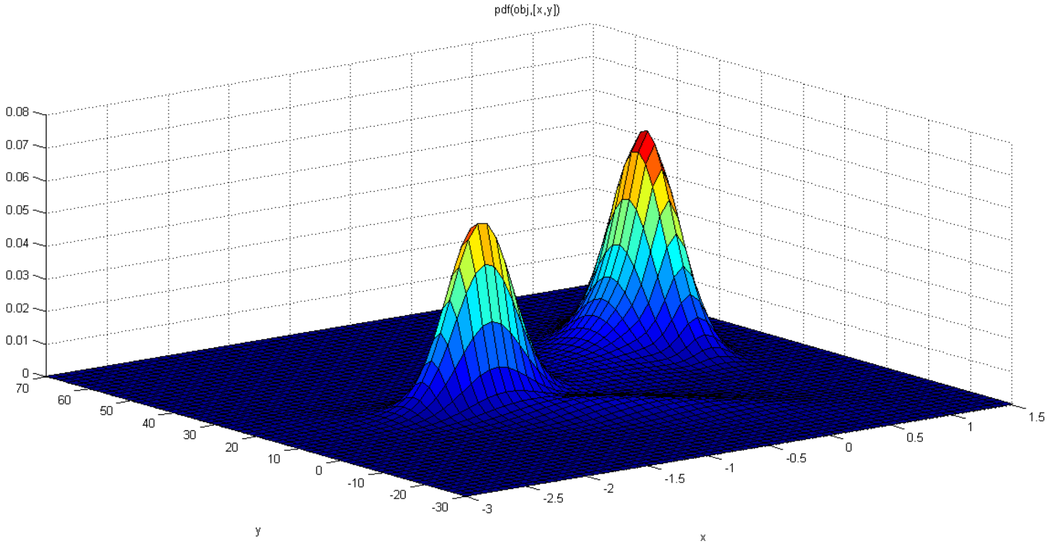


Figure 12: Surface diagram of first two dimensions from a feature vector after training

The GMM system is simple to train because it requires neither an orthographic transcription nor phonetic labeling of the training speech.

4.2 Maximum Likelihood Parameter Estimation

For a given training vectors and a GMM configuration, we have to estimate the parameters of the GMM, λ, for the best matches for the distribution of the training feature vectors. The most popular and well-known method is maximum likelihood (ML) estimation.

During recognition, an unknown speech utterance is classified by first converting the digitized waveform to feature vector, X, comprising of observations $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_T$ and then by calculating the average log likelihood that the dialect model produced the unknown speech utterance. The log likelihood L is defined as [21],

$$L(X|\lambda_d) = \frac{1}{T}\sum_{t=1}^{T} \log p(\bar{x}_t|\lambda_d) \qquad (14)$$

Where, λ_d is the model for the corresponding dialect d and T is the total observation. The maximum-likelihood classifier hypothesizes has the dialect of the unknown utterance can be calculated as,

$$H = \ arg \max_d L(X|\lambda_d) \qquad (15)$$

Chapter 5

Experimental Result & Analysis

Gaussian mixture modeling has been implemented for our system. We have collected voice samples for training and testing phases and this process come to an end with some experimental results.

5.1 Sample Collection

Sample collection occupies a major part of our research. It is actually a challenging task. There are many dialects in Bangladesh as well as it was very time consuming for us to work with all in a short period. Initially we train our system with Borishali, Noakhali, Sylheti ,Chittagonian and Chapai dialects because this dialects have more variance. As accuracy rate is about 75%, so we can train the system with other dialects.

Smart Voice Recorder has been used for recording the voice samples at 16 KHz sample rate and 16bits quantization. About 30 samples of each dialect have been collected for training and testing the interface. Other than that, many other samples that we collected could not be used due to lack of perfection.

Duration of each voice sample is 3 minutes. 3 minutes appear really large period while recording and we face many difficulties. It is incredibly hard for an urban people to speak frequently in their regional language. We try our level best to take the samples in a noise free environment, but ensuring it in our locality is rare. In addition gathering people from different dialects and ensuring their fluency in their particular regional language is quite difficult. We have to explain each of the sample provider our purpose of collecting their samples and make them convinced. Delay occurs due to unavailability of suitable place to conduct the requirement while the recordings. Furthermore we have to consider their emotion while speaking. In reality, it is not easy to speak continuously for three minutes, as a result it take time and patience for the recordings.

Many researchers cannot proceed with their research work since there is no structured Bengali database. We also feel the importance of building a database of Bengali language. Our data collection has not yet finished. We will include the remaining dialects and create a database for the researchers.

5.2 Result Discussion

In this works, the regional language recognition system was developed using Gaussian Mixture Model approach. The dialect identification experiment was trailed using 5 regions of Bangladesh. In our experiments, around 150 test speakers of different dialect are taken for training and testing phase. Worth of note is GMM provides maximum accuracy rate as found in [2]. Therefore our approaches were taken using GMM system. Table 5.2 shows all the experimental results given as the percentage of dialects correctly identified. The result of regions, we worked with is shown below in a table.

Dialect	Recognition Rate %	Error %
Chittagonian	**60.61**	**39.39**
Sylheti	**80**	**20**
Noakhali	**72.41**	**27.59**
Borishali	**83.33**	**16.67**
Chapai	**90**	**10**

Table 1: Experimental Result of Our System Configuration

Experimental result has diversity due to the variation of dialects. Computation speed up approach of GMM is also responsible for the diversity. During the training phase, the system takes the speech samples for single dialect and converts them into feature vectors. During the recognition, an unknown sample is compared to each of the GMMs [1].

So it depends on the probability of generating a new point on the defined model which is already built for detecting a specific dialect.

GMMs of the dialects can be overlapped. If testing sample gets plotted into the overlapped area between two different GMM of dialects, then system can provide a wrong recognition.

We are experimenting with different dialects of a same language where most of the words are same. Moreover language maintains a specific pattern, alignment and grammatical order. Bengali language has many similarities but variations in dialects. As a result error occurs. Chapai Nobabgonj dialect detection shows comparable performance which is maximum 100 percent.

Chapter 6

Conclusion and Future Work

Several research works have been performed to detect speeches, dialects and languages of different region throughout the world. But no initiative has been taken for Bangladeshi regional dialect. A method has been proposed to detect a Bangladeshi different dialect which utilizes Mel-Frequency Cepstral Coefficients (MFCC), its Deltas and Delta-Deltas as main features and Gaussian Mixture Models (GMM) to classify characteristics of a specific dialect. The methodology is explained elaborately in above chapters with necessary diagrams.

Initially the scheme is tested on 5 databases of Barishal, Noakhali, Sylhet, Chittagong and Chapai Nobabgonj regions. As experimental result show that GMM adaptation gives comparable good performance and well accuracy rate, consequently we can train the system with other dialects.

Our core contribution of thesis is sample collection for training and recognition. Many researchers can not complete their thesis since there is lack of Bengali database. So we have decided to build a database for them and continue our sample collection. Experimental result has diversity and errors. We will also work on improving its recognition rate in future.

In future, the system can be expanded to discriminate among more dialects. We are also interested to implement Artificial Neural Network (ANN) based classifier to detect dialects and compare it with GMM based approach.

References

[1] E. Wong, J Pelecanos, S. Myers and S. Sridharan. “Language Identification Using Gaussian Mixture Model Analysis,” Speech Research Lab, RCSAVT, School of Electrical and Electronic Systems Engineering, Queensland University of Technology.

[2] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk. “Speech Recognition using MFCC,” International Conference on Computer Graphics, Simulation and Modeling , 2012, page(s): 135-138.

[3] William Campbell, Terry Gleason “Advanced Language Recognition using Cepstral and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation”. In Proc. IEEE Odyssey, 2006, page(s): 1-8.

[4] H. Suo, M. Li, P. Lu, and Y. Yan, “Automatic language identification with discriminative language characterization based on svm”, IEICE-Transactions on Info and Systems, Volume E91-D, Number 3 , Pp. 567-575,2008.

[5] T. Peng, W., and B. Li, "SVM-UBM based automatic language identification using a vowel-guided segmentation", Third International Conference on Natural Computation (ICNC 2007), ICNC, pp. 310-314, 2007.

[6] P.A. Torres-Carrasquillo, D.A. Reynolds, and J.R. Deller, "Language identification using Gaussian mixture model tokenization", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02), vol.1, no., pp. I-757-I-760 vol.1, 2002.

[7] A.D. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted gaussian mixture models”, Digital Signal Processing, Vol. 10, 19–41 (2000).

[8] R. Córdoba, L.F. D‟Haro, R. San-Segundo, J. Macías-Guarasa, F. Fernández, and J.C. Plaza, “A multiple-Gaussian classifier for language identification using acoustic information and PPRLM scores”, IV Jornadas en Tecnologia del Habla, 2006, Pp. 45-48.

[9] P.A. Torres-Carrasquillo, T.P. Gleason, and D.A. Reynolds, “Dialect identification using gaussian mixture models”, In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 297-300, 31 May - 3 June 2004.

[10] Shrawankar, Urmila, and Vilas M. Thakare. "Techniques for feature extraction in speech recognition system: A comparative study." arXiv preprint arXiv:1305.1145 (2013).

[11] G. Gaurav, D. Deiv, G. Sharma and M. Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi," Journal of Signal and Information Processing, Vol. 3 No. 3, 2012, pp. 394-401. doi:10.4236/jsip.2012.33052.

[12] Muda, Lindasalwa, Mumtaj Begam, and I. Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[13] IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[14] D. Reynolds, R.Rose, “Robust text-independent speaker identification using Gaussian Mixture Models”, IEEE Trans. Speech Audio Process, vol 3, no.1,pp 72-83, Jan.1995.

[15] N. Malayath, H. Hermansky, S. Kajarekar, B. Yegananarayan, “Data–driven temporal filters and alternatives to GMM in speaker verification”, Digital Signal Processing, 55-74,2000.

[16] D. Reynolds, “Gaussian Mixture Models*”, MIT Lincoln Laboratory,244 wood St. Lexinton, MA 02140,USA.

[17] A. Fazel and S. Chakrabartty, “An overview of Statistical Pattern Recognition Techniques for Speaker Verification”, IEEE Circuits and System Magazine, 62-81, 2011.

[18] L. Riek, W. Mistrerra, and D. Morgan, "Expetimenrs in Language Idenrification," Technical Report SPCOT-91-002 (Lockheed-Sandets, Nashua, NH, Dec. 1991).

[19] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-Independenr, Text Independent Language Identification by HMM," ICSLP '92 Proc. 2, Banff,Alberta, Canada, 12-16Oct. 1992, p. 1011.

[20] M.A Zissman, "Auromatic Language Idenrification Using Gaussian Mixture and Hidden Markov Models," ICASSP '93 Proc. 2, Minneapolis, 27-30Apr. 1993, p. 399.

[21] Wong, E., et al. "Language identification using efficient Gaussian mixture model analysis." Australian International Conference on Speech Science and Technology. Vol. 4. 2000.

[22] Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Comm. COM-28, 84 (1980).

[23] R.G. Leonard and G.R. Doddingron, "Auromatic Language Idenrification," Technical Report RADC-TR-74-200/TI347650 (RADC/Texas Insrrumenrs, Dallas, Aug. 1974).

[24] R.G. Leonard and G.R. Doddingron, "Automatic Classification of Languages," Technical Report RADC-TR-75-264 (RADC/Texas Instruments, Dallas, Oct. 1975).