January 2012

Speaker Recognition Using Shifted MFCC

Rishiraj Mukherjee
University of South Florida, rishiraj@mail.usf.edu

Speaker Recognition Using Shifted MFCC

by

Rishiraj Mukherjee

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Electrical Engineering
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Ravi Sankar, Ph.D.
Wilfrido Moreno, Ph.D.
Paris Wiley, Ph.D.

Date of Approval:
March 19, 2012

Keywords: Fourier Transform, Accent Modeling, Speech Processing,
MFCC,Gaussian Mixture Model

DEDICATION

Dedicated to my parents who sacrificed their today for our better tomorrow and my mentors who have guided me throughout my research and helped me to improve professionally and personally.

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the guidance and support of my thesis advisor, Dr. Ravi Sankar, whose insightful comments and explanations have taught me a great deal about speech and research in general. I am also grateful to Dr. Wilfrido Moreno and Dr. Paris Wiley for serving on my committee. I would also like to thank iCONS group members, especially Tanmoy Islam for their valuable comments on this work. Finally I would like to thank my mother Chandrabali my father Basab and my sister Bindi, for their encouragement, support and love.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Speaker Recognition is the art of recognizing a speaker from a given database using speech as the only input. In this thesis we will be discussing a novel approach to detect speakers. Here we will introduce the concept of shifted MFCC to add improvement over the performance from previous work which has shown quite a decent amount of accuracy of about 95% at best. We will be talking about adding different parameters which also contributed in improving the efficiency of speaker recognition. Also we will be testing our algorithm on Text dependent speech data and Text Independent speech data. Our technique was evaluated on TIDIGIT – database. In order to further increase the speaker recognition rate at lower FARs, we combined accent information added with pitch and higher order formants. The possible application areas for the work done here is in any access control entry system or now a day's a lot of smart phones, laptops, operating systems etc have Also, in homeland security applications; speaker accent will play a critical role in the evaluation of biometric systems since users will be international in nature. So incorporating accent information into the speaker recognition/verification system is a key component that our study focused on. The accent incorporation method and Shifted MFCC techniques discussed in this work can also be applied to any other speaker recognition systems.

CHAPTER 1

INTRODUCTION

1.1 Background

There have been a lot of recent developments in recent times, in the era of internet and computers: the digital computer, and the era of industrial automation, everything is done and kept track of through computers. From banking to automated nuclear power plants, we live in a generation when we have made amazing progress in terms of mathematics, control theory, fabrication All these developments have contributed to advancement of technology. But along with advancement of technologies, security threats have increased in various realms such as information, airport, home, international, and national securities. According to [1], Identity thefts cost US $56.6 billion per year. According to the same paper, experts say many incidents go undetected or unreported. Due to the increased level of security threats and fraudulent transactions, the need for reliable user authentication has increased and hence biometric security systems have emerged Speaker recognition is the task of automatically recognizing people from their speech signals. This technique makes it possible to use uttered speech to verify the speaker's identity and control access to secure services, i.e., online transactions, database access services, information services, security control for confidential information areas, remote access to computers etc. Text dependent speaker recognition follows the technique of detecting speakers based on the text. i.e all the speakers will be saying the same thing and the goal

is to distinguish the speakers. Text dependent speaker recognition was chosen because a lot of the access based security systems use speech as a way of blocking unwanted individuals. The main goal was to increase the accuracy of the text-dependent speaker recognition performance. Now a days there is a lot of talk going about creating hardware modules which can take care of recognition of speech or face or other patterns as iris recognition and the list goes on. The primary reason behind this is the fact that we all want to automate things like speech recognition, which reduces manpower in many areas and also because of the vastness of data available these days, it is humanely impossible to recognize patterns within these data's. Speech is more of a behavioral part of the human being as it deals with the vocal tract of a human being. The reason behind going for algorithms like MFCC ,GMM is the fact that they are very popular and also they produce accurate results. The importance of the field of biometrics is going stronger and even a lot of the popular operating systems are incorporating these features and not only Operating systems, we are seeing several popular social websites are using these features to classify different persons. The reason behind choosing speech the biometric system I am talking about is that speech is the most intuitive way by which we recognize a person. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear" spectrum-of-a-spectrum").

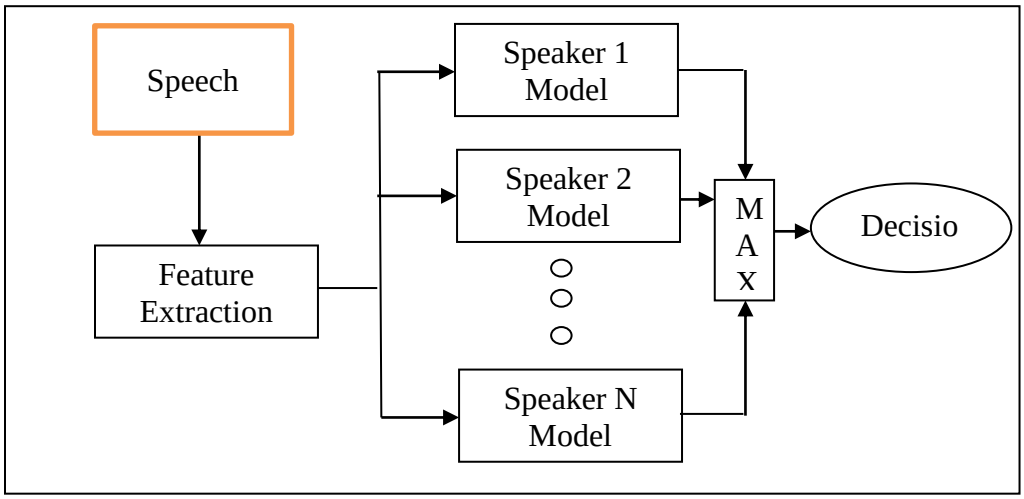


Figure 1. Speaker Identification System

A typical speaker recognition system is made up of two components; feature extraction and classification. Speaker recognition (SR) can be divided into *speaker identification* and *speaker verification*. Speaker identification system determines who amongst a closed set of known speakers is providing the given utterance as depicted by the block diagram in Figure 1. Speaker specific features are extracted from the speech data, and compared with speaker models created from voice templates previously enrolled. The model with which the features match the most is selected as the legitimate speaker. In most cases, the model generates a likelihood score and the model that generates the maximum likelihood score is selected.

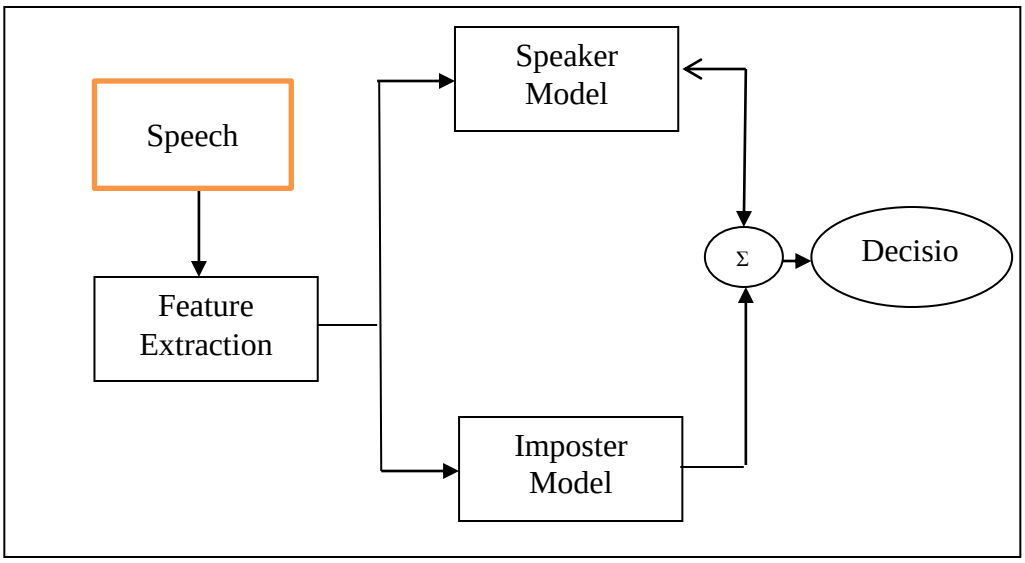


Figure 2. Speaker Verification System

On the other hand, speaker verification system as depicted by the block diagram in Figure 2,accepts or rejects the identity claim of a speaker. Features are extracted from speech data and compared with the legitimate speaker model as well as an imposter speaker model, which are created from previously enrolled data. The likelihood score generated from the speaker model is subtracted from the imposter model. If the resultant score is greater than a threshold value, then the speaker is accepted as a legitimate speaker. In either case, it is expected that the persons using these systems are already enrolled. Besides these systems can be text dependent or text independent. Text dependent system uses a fixed phrase for training and testing a speaker. On the contrary, text independent system does not use a fixed phrase for training and testing purposes. In addition to

security, speaker recognition has various applications and is rapidly increasing. Some of the areas where speaker recognition can be applied are [3]:

1) Access Control:

Secure physical locations as well as confidential computer databases can be accessed through one's voice. Access can also be given to private and restricted websites.

2) Online Transactions:

In addition to a pass phrase to access bank information or to purchase an item over the phone, one's speech signal can be used as an extra-layer of security.

3) Law Enforcement:

Speaker recognition systems can be used to provide additional information for forensic analysis. Inmate roll-call monitoring can be done automatically at prison.

4) Speech Data Management:

Voicemail services, audio mining applications, and annotation of recorded or live meetings can use speaker recognition to label speakers automatically.

5) Multimedia and personalization:

Sound tracks and music can be automatically labeled with singer and track information. Websites and computers can be customized according to the person using the service.

1.2 The Problem

Even though speaker recognition systems have been researched over several decades and have numerous applications, they still cannot match the performance of a human recognition system [4] and as well as not reliable enough to be considered as a standalone

security system. Although speaker verification is being used in many commercial applications, speaker identification cannot be applied effectively for the same purpose. The performance of speaker recognition systems degrade especially under different operating conditions. Speaker recognition system performance is measured using various metrics such as recognition or acceptance rate and rejection rate. Recognition rate deals with the number of genuine speakers correctly identified, whereas rejection rate corresponds to the number of imposters (people falsifying other's identity) being rejected. Along with these performance metrics there are some performance measures and tradeoffs one need to consider while designing speaker recognition systems. Some of the performance measures generally used in the evaluation of these systems include: false acceptance rate (FAR) - the rate at which an imposter is accepted as a legitimate speaker, true acceptance rate (TAR) - the rate at which a legitimate speaker is accepted, and false rejection rate (FRR) -the rate at which a legitimate speaker is rejected(FRR=1-TAR).

There is a trade-off between FARs and TARs, as well as between FARs and FRRs. Intuitively, as the false acceptance rate is increased, more speakers are accepted, and hence true acceptance rate rises as well. But the chances of an imposter accessing the restricted services also increase; hence a good speaker recognition system needs to deliver performance even when the FAR threshold is lowered. The main problem in speaker recognition is, poor TARs at lower FARs, as well as high FRRs. The performance of a speaker recognition system [3]for three different datasets is shown in Figure 3. Here, error (%) which is equivalent to FRR (%) has been used to measure performance. The TIMIT dataset consists of clean speech from 630 speakers. As the

dataset is clean we can see that the error is almost zero, even though the number of people is increased from 10 to 600. For NTIMIT, speech was acquired through telephone channels and the performance degraded drastically as the speaker size was increased. At about 400 speakers we can see that the error is 35%, which means a recognition rate of 65%. We can see the similar trend for SWBI dataset, where speech was also acquired through telephone

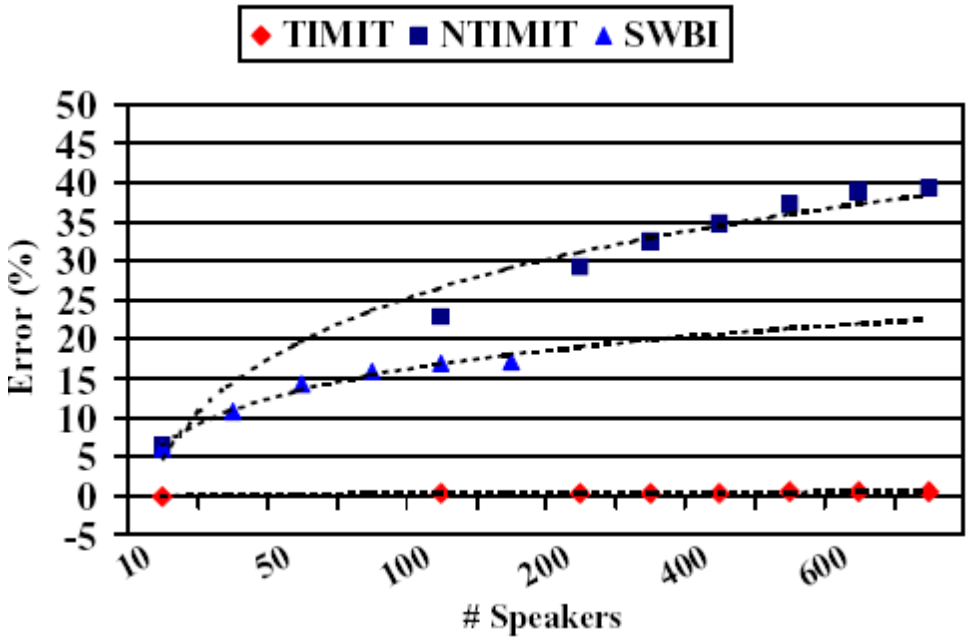


Figure 3. Current Speaker Recognition Performance Over Various Datasets[3]

channel. But the performance for SWBI is not as low as TIMIT, which indicates that various other factors other than the type of acquisition influence the recognition rate. It depends on the recording quality (environmental noise due to recording conditions and noise introduced by the speakers such as lip smacks) and the channel quality. Hence it is hard to generalize the performance of an SR system on a single dataset. From Figure 3, we can see that the recognition rate degrades as the channel noise increases and also

when the number of speakers increases. Another evaluation of current voice recognition systems (Figure 4) conducted by the UK BWG (Biometric Working Group)shows that about 95% recognition can be achieved at an FAR of 1%[5]. The dataset consisted of about 200 speakers and voice was recorded in a quiet office room environment.

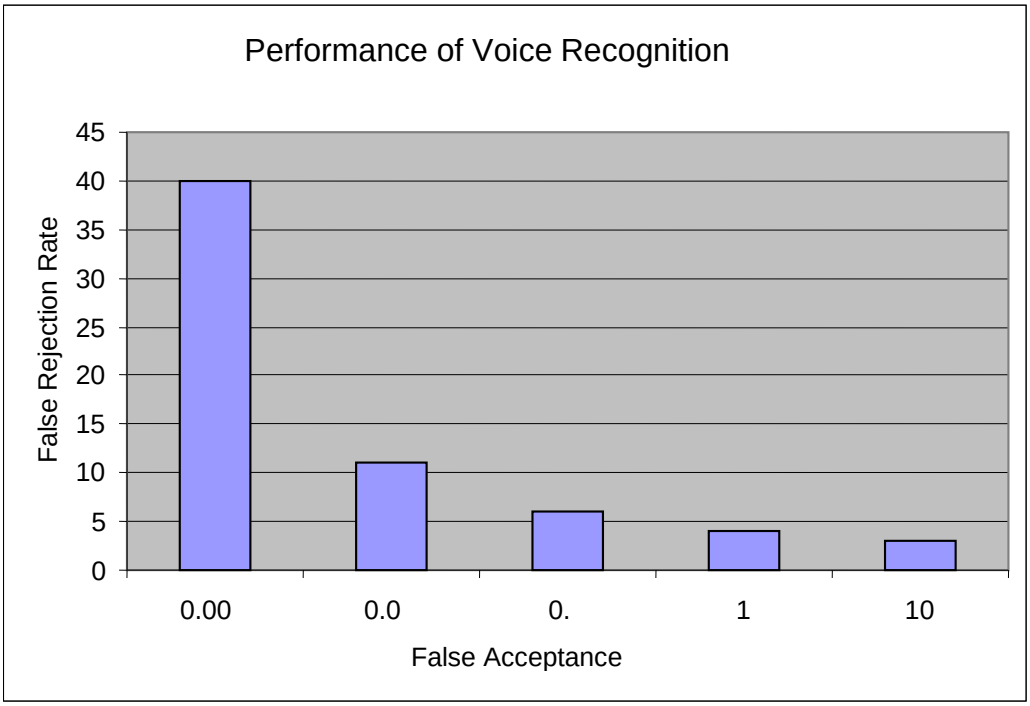


Figure 4. Current Speaker Recognition Performance Reported by UK BWG[5]

On the whole, we can see that speaker recognition performance in a real world noisy scenario cannot provide a high level of confidence. Speaker recognition systems can be considered reliable for both defense and commercial purposes, only if a promising recognition rate is delivered at low FARs for realistic datasets.

1.3 Motivation

In this thesis, an effort has been made to deal with the problem, i.e. to achieve high TAR at lower FARs even in realistic noisy conditions, by enhancing recognition performance with the help of intra-modal fusion and accent modeling. The motivation behind the thesis can be explained by answering the three questions – why enhance speaker recognition, why intra-modal fusion and why combine accent information. In case of speaker recognition, obtaining a person's voice is non-invasive when compared to other biometrics, for example capture of iris information. With very little additional hardware it is relatively easier to acquire this biometric data. Recognition can be achieved even from long distance via telephones. In addition surveillance, counter-terrorism and homeland security department can collect voice data from telephone conversation without having to access to any other biometric dataset. In this type of scenario it would be beneficial if the confidence level of authentication is high. Previous research works in biometrics have shown recognition performance improvements by fusing scores from multiple modalities such as face, voice, and fingerprint[6], [7], [8].However multi-modal systems have some limitations, i.e., cost of implementation, availability of dataset, etc. On the other hand, by fusing two algorithms for the same modality (intra-modal fusion), it has been observed in [8], that performance can be similar to inter-modal systems when realistic noisy datasets are used .Intra-modal fusion reduces complexity and cost of implementation when compared to various other biometrics, such as fingerprint, face, iris etc. Various additional hardware and data is required for acquiring different biometrics of the same person.

Finally, speech is the most developed form of communication between humans. Humans rely on several other types of information embedded within a speech signal, other than voice alone. One of the higher levels of information that humans use is accent. Also, incorporation of accent information provides us with a narrower search tool for the legitimate speaker in huge datasets. In an international dataset, we can search within a pool of dataset, where speakers belong to the same accent group as the legitimate speaker. Homeland security, banks and many other realistic entities, deal with users who are international in nature. Hence incorporation of accent is a key for our speaker recognition model.

1.4 Thesis Goals and Outline

The main goal in this thesis is to enhance speaker recognition system performance at lower FARs with the help of an accent classification system, even when evaluated on a realistic noisy dataset. Thus the final enhanced recognition score is achieved. Our system consists of three parts HF system, AC system and the score modifier (SM) algorithm. The HF speaker recognition system[9]is made up of score-level fusion of AHS[10] and GMM [11]models, which takes enrolled and test speech data as inputs and generates a score as an output, which is a matrix when a number of test speech inputs are provided. The accent classification system is made up of a fusion of Gaussian Mixture Model (GMM)[12], and) [13], as well asa reference accent database. It accepts enrolled and test speech inputs and generates an accent score and an accent class as the outputs for each test data. The SM algorithm, a critical part of the proposed system, makes mathematical modifications to the resultant HF score matrix controlled by the outputs of the accent

classification system. The final enhanced recognition scores are generated after the modifications are made to the HF scores by the score modifier. Feature extraction is an internal block within both the HF system as well as the accent classification (AC) system.

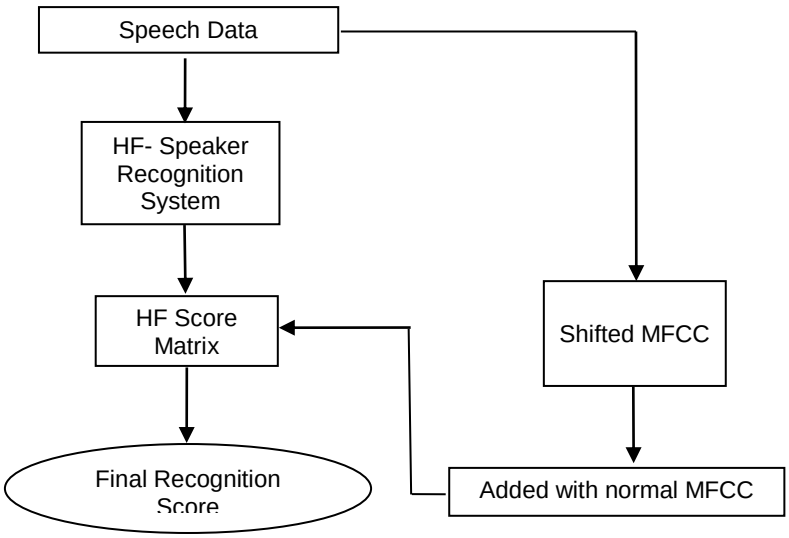


Figure 5. Algorithm Outline

The rest of the thesis is organized as follows. In the next sections each segment of the HFA system is described thoroughly in the next chapters. The hybrid fusion speaker recognition is explained in Chapter 2, which consists of background information of speech, feature extraction, speaker model creation and the fusion technique used to fuse the speaker recognition models. In Chapter 3, the accent classification system is described, along with past research work in accent classification, accent feature, and the formulation of accent classifier. In Chapter 4, the combination of speaker and accent models is investigated and its effects are studied. Chapter 5 describes the datasets and shows the results and performances of hybrid fusion, accent classification and the

complete system. Finally, Chapter 6 contains the conclusions and recommendation for future research.

CHAPTER 2

SPEAKER RECOGNITION SYSTEM

2.1 Overview of Past Research

Pruzansky at bell labs in 1960, was one of the first ones to research on speaker recognition, where he used filter banks and correlated two digital spectrograms for a similarity measure[14]. P. D. Bricker and his colleagues experimented on text-independent speaker recognition using averaged auto-correlation[15]. B. S. Atal studied the use of time domain methods for text-dependent speaker recognition[16]. Texas Instruments came up with the first fully automatic speaker verification system in the 1970's. J. M. Naik and his colleagues researched the usage of HMM techniques instead of template matching for text-dependent speaker recognition[17]. In [18], text independent speaker identification, was studied based on a segmental approach and mel-frequency cepstral coefficients were used as features. Final decision and outlier rejection was based on a confidence measure. T. Matsui and S. Furui investigated vector quantization (VQ) and HMM techniques to make speaker recognition more robust[19]. Use of Gaussian mixture models (GMM) for text-independent speaker recognition was successfully investigated by D. A. Reynolds and R. Rose[12]. In the 2000's, research focused on adding higher level information to speaker recognition systems, to makethe systems more robust and to increase the confidence level of speaker recognition

systems. G. Doddingtonused ideolectic features of speech such as word unigrams and bigrams to characterize a certain speaker[20]. Evaluation was performed on the NIST extended data task, which consisted of telephone quality long duration speech conversation from 400 speakers. A miss rate of 40%was observed at an FAR of 1%. In 2003, A. G. Adami, used temporal trajectories of fundamental frequencies and short term energies to segment and label speech, which were then used to model a speaker with the help of an n-gram model[21]. The same NIST extended dataset was used and similar performance as in [20] was observed. In 2003, D. A. Reynolds and his colleagues used high level information such as pronunciation models, prosodic dynamics, pitch and duration features, phone streams and conversational interactions, which were fused and modeled using an MLP to fuse n-grams, HMMs and GMMs[22]. The same NIST dataset was used for evaluation and a 98% TAR was observed at 0.2% FAR. Also in 2006, a multi-lingual NIST dataset consisting of 310 speakers was used for cross lingual speaker identification. Several speaker features derived from short time acoustics, pitch, duration, prosodic behavior, phoneme and phone usage were modeled using GMMs, SVMs and N-Grams[23]. The several modeling systems used in this work, were fused using a multi layer perceptron (MLP). A Recognition rate of 60% at an FAR of 0.2% has been reported. In [24], mel-frequency cepstral coefficients (MFCC) were modeled using phonetically structured GMMs and speaker adaptive modeling. This method was evaluated on YOHO consisting of clean speech from 138 speakers and Mercury dataset consisting of telephone quality speech from 38 speakers. An error rate of 0.25% on YOHO and 18.3% on Mercury were observed. In [25], MFCCs and their first order derivatives were used as features and an MLP fusion of GMM-UBM system and speaker

adaptive ASR system were used to model these features. When evaluated on the Mercury and Orion datasets consisting of 44 speakers in total, a miss rate of 7.3% has been reported. In [26], a 35 speaker NTT dataset, was used for evaluating a fusion of a GMM system and a syllable based HMM adapted by MAP system. MFCCs were used as features and a 99% speaker identification has been reported. In[27], SRI prosody database and NIST2001 extended data task were used for evaluation. Though this paper was not explicitly considering accent classification, it used a smoothed fundamental frequency contour (f0) at different time scales as the features. The f0 contour was then converted to wavelets by wavelet analysis. The output distribution was then compacted and used to train a bigram for universal background models (UBM) using a first order Markov chain. The log likelihood scores of the different time scales were then fused to obtain the final score. The results indicate an 8% EER on a DET curve for 2 utterance test segments and it degrades to 18% when 20 test utterance segments were used. The NIST 2001 extended data task consisting of 482 speakers was used for evaluation. In [28],exclusive accent classification was not performed, but formant frequencies were used for speaker recognition. Formant trajectories and gender were used as features and a feed forward neural network was used for classification. A 6 speaker dataset was created from the TIMIT database. An average 6.6% misclassification rate was observed for the six speakers.In this thesis, we focused on an intra-modal speaker recognition system, to achieve similar performance enhancement observed in [6], [7]. However, we used two complementary voice recognition systems and fused their scores to have a better performing system. Similar work has been done in [24], [25] and [26], where scores from two recognition systems were fused and one of the recognition algorithm was a variant of

Gaussian Mixture Model (GMM) [24] and the other being a speaker adapted HMM [26]. But, there are a number of factors that differentiate this work from those described in [24], [25] and [26]: Database size, data collection method, and the location of the data collected (indoor and outdoor dataset). In [25] and [26], a small dataset, population of 44 and 35 respectively, was used. We, on the other hand, conducted our experiment on two comparatively larger indoor and outdoor datasets. There has been a great deal of research towards improving speaker recognition rate by adding supra-segmental, higher level information and some accent related features like pronunciation models and prosodic information [21], [22], [27], [28]. But the effect of incorporating the outcome of an accent modeling/classifying system into a speaker recognition system has not been studied so far. Also in [21] and [22], performance of the systems was good, but at the cost of complexity due to the utilization of several classifiers with various levels of information fusion. But the system developed in this thesis has a reduced level of complexity compared to these higher level information fusion systems.

2.2 Algorithm Outline

Figure 6 shows the flow chart of our proposed method. We used same person's voice data from each dataset to extract features. Arithmetic Harmonic Sphericity (AHS) is used to generate a similarity score between the enrolled feature and the test feature. A Gaussian Mixture Model is created from enrolled features and an GMM likelihood score is generated for each test feature. The AHS and GMM likelihood scores form a matrix as in a training and testing scenario, there are a number of speakers being enrolled and tested.

These score matrices are then fused using a linear weighted hybrid fusion methodology to generate intra-modal enhanced scores. The features used and the speaker models used to generate likelihood scores, as well as the fusion methodology is explained next.

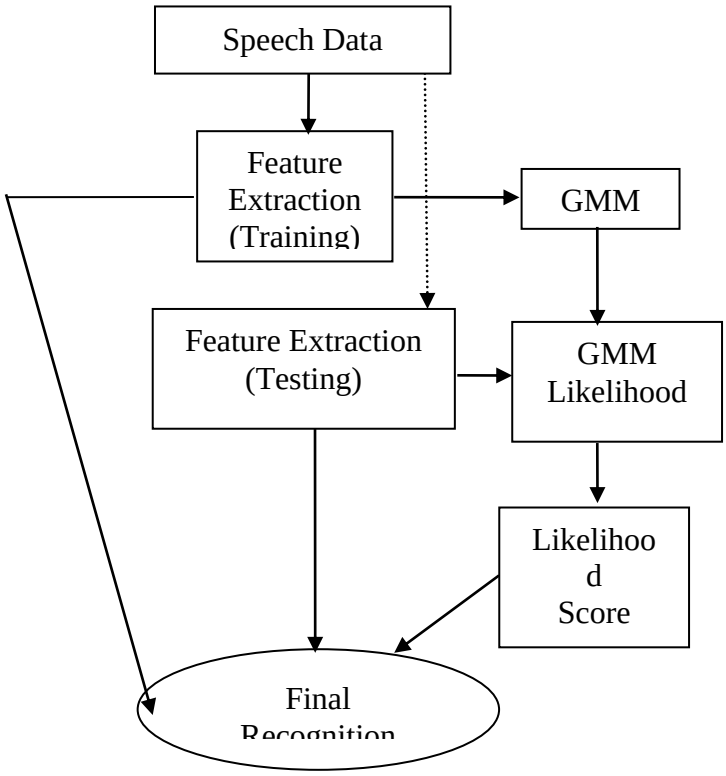


Figure 6. Shifted MFCC Modeling

Here initially after the speech data is first passed through a low pass filter to remove noise , it is used to extract the formants and pitch which after added with the shifted MFCC, explained in the section below is used for GMM to model. The next step is to get a score by using a distance metric from the GMM from[1] and thus we have a model which can distinguish speakers.

2.3.Speech Science and Feature Extraction

Before feature extraction is described it is necessary to understand the type of signal we are dealing with. The rest of this section explains the characteristics of speech signals, various pre-processing steps involved in feature extraction and finally the feature itself.

2.4. Speech Signal Characteristics and Pre-Processing

Speech is produced when a speaker produces a speech signal in the form of a sound pressure wave that travels from the speaker's mouth to a listener's ears. Speech signals are composed of a sequence of sounds that serve as a symbolic representation for a thought that the speaker

wishes to convey to the listener. The arrangement of these sounds is governed by a set of rules which is called language[29].

A speech signal must be sampled in order to make this data available to a digital system as natural speech is analog in nature. Speech can be represented as a 2- Dimensional vector, consisting of number of samples and the corresponding amplitude. For example the digit "Six" when plotted would be as shown in Figure 7. This speech signal is phonetically represented as /s/ /i/ /k/ /s/, where s, i, k, s are phonemes (smallest unit of sound) of the speech signal.

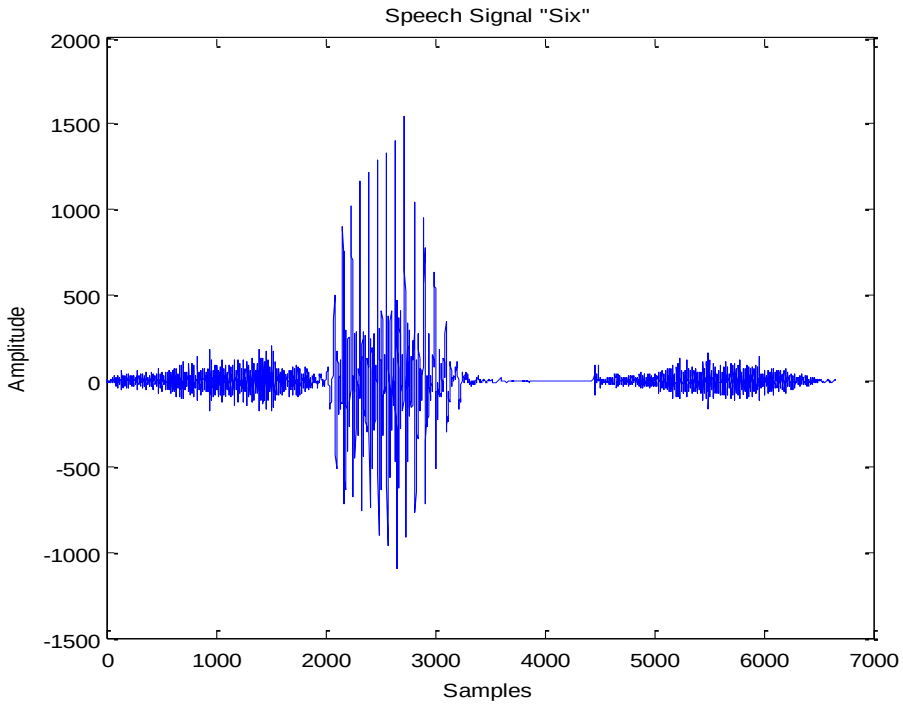


Figure 7. Example of a Speech Signal

Speech sounds can be classified into voiced, unvoiced, mixed and silence segments. Voiced sounds have higher energy levels and are periodic in nature whereas unvoiced sounds are lower energy sounds and are generally non-periodic in nature. Mixed sounds have both the features, but are mostly dominated by periodic nature. In order to distinguish speech of one speaker from the speech of another, we must use a feature of the speech signal which characterizes a particular speaker, which is a rather complicated task when a speech signal is considered because of the time - varying property of a speech signal. Hence speech, which is quazi-stationary has to be divided into sound segments that possess similar acoustic properties over short periods of time before features can be extracted.

1) Framing

As discussed in the previous section, we need to process speech data in small chunks called frames rather than the whole data itself. Typically, a frame of 20 - 30 ms could be considered for the short time Time-Invariant property of speech signal. Framing can be classified as non-overlapping and overlapping frames. An example of a frame extracted from the speech data “six” is shown in Figure 8.It can be noted that the signal is periodic in nature, because the extracted frame consists of voiced sound /i/.

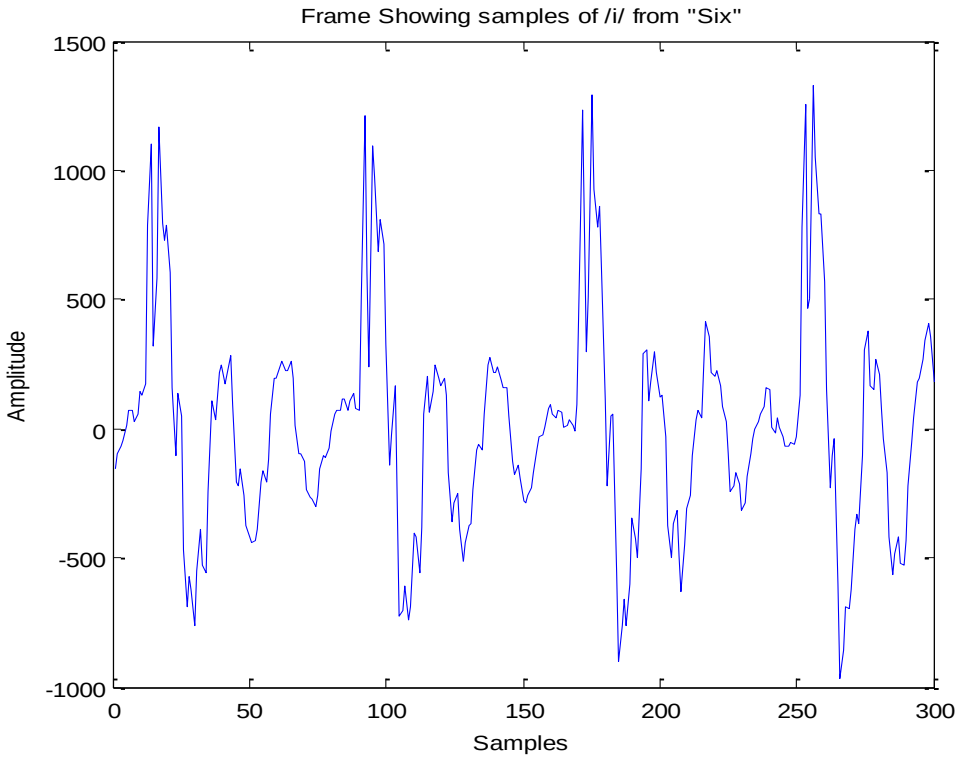


Figure 8. Example of Framing

2) Windowing

A window is a real finite length sequence used to select a desired frame of the original signal by a simple multiplication process and this process is a called windowing. A frame of the signal x(n) of length N (same as the duration of the window) ending at time m, say f (n;m) is obtained as f (n;m) = x(n)w(m-n). A meaningful window should be able to preserve spectral information as well as have less spectral leakage.

A rectangular window (high spectral leakage) is defined as,

$$w(n) = \begin{cases} 1, & \text{n=0,1,.....N-1} \\ 0, & \text{n otherwise} \end{cases} \quad (1)$$

A hamming window (low spectral leakage) is defined as,

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/\text{N-1}), & \text{n=0,1,.....N-1} \\ 0, & \text{n otherwise} \end{cases} \quad (2)$$

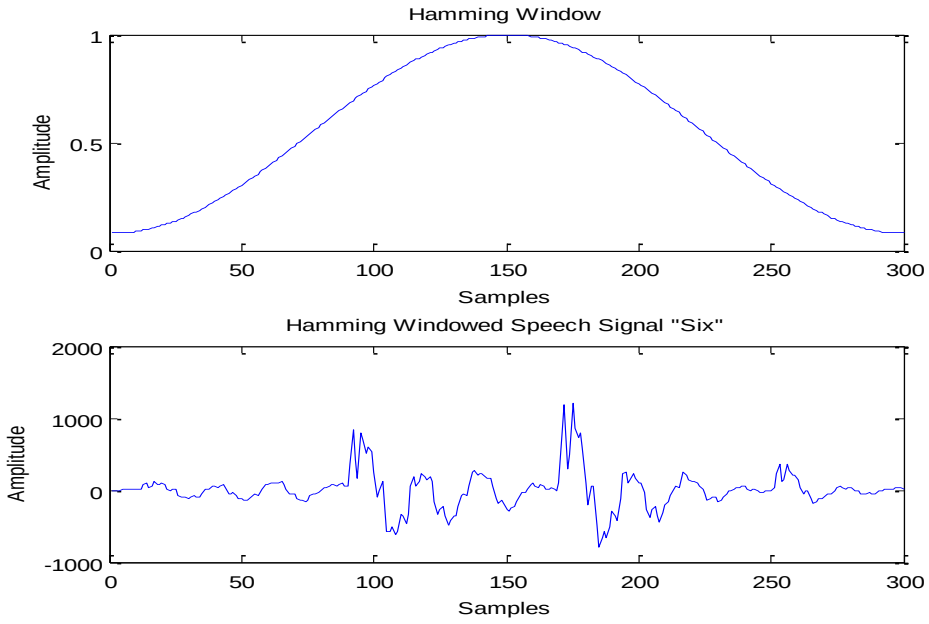


Figure 9. Example of Windowing

As seen in the Figure 9, the middle portion of the signal is enhanced whereas the beginning and the end samples are attenuated as a result of using a Hamming window, but it is useful as preserves spectral properties of the signal.

3) Fast Fourier Transform

Fast Fourier Transform (FFT) is a name collectively given to several classes of fast algorithms for computing the Discrete Fourier Transform (DFT). The Discrete Fourier Transform provides a mapping between the sequence, say x (n), n=0, 1, 2……, N-1 and a discrete set of frequency domain samples, given by

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)kn}, & \text{k=0,1,.....N-1} \\ 0, & \text{k otherwise} \end{cases} \quad (3)$$

The inverse DFT (IDFT) is given by

$$x(n) = \begin{cases} \frac{1}{N} \sum_{n=0}^{N-1} X(k) e^{j(2\pi/N)kn}, & \text{n=0,1,.....N-1} \\ 0, & \text{n otherwise} \end{cases} \quad (4)$$

Where, the IDFT is used map the frequency domain samples back to time domain samples. The DFT is always is periodic in nature, where k varies from 1 to N, where N is the size of the DFT. The Figure 10 shows a 512-Point Fast Fourier Transform for the speech data “Six”.

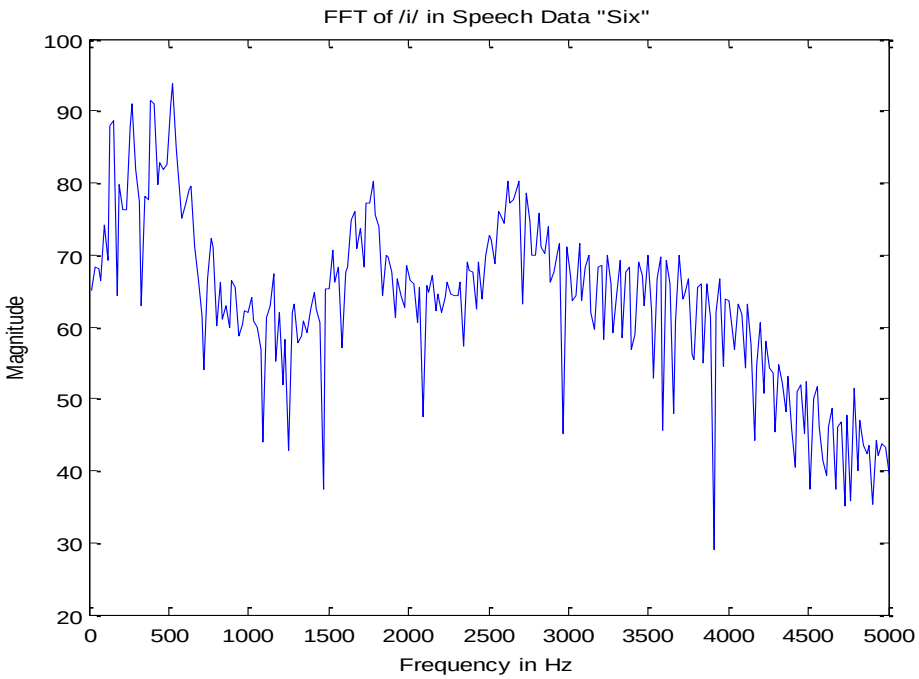


Figure 10. Example of FFT

4) Cepstrum Domain

Speech is composed of an excitation sequence convolved with the impulse response of the vocal system model. It is often desirable to eliminate one of the components so that the other may be used in a recognition algorithm. Cepstrum is a common transform, which can be used to separate the excitation signal (which contains the phones and the pitch) and the transfer function (which contains the voice quality). These two portions are convolved in the time domain, but convolution in time domain becomes multiplication in frequency domain, which could be represented as,

$$X(\omega) = G(\omega)H(\omega) \tag{5}$$

When a log of the magnitude of both sides of the transform is taken,

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \tag{6}$$

Taking IDFT on both sides of the above equation, introduces us to a term called 'Quefrency', which is the x-axis of the cepstrum domain.

$$IDFT(\log|X(\omega)|) = IDFT(\log|G(\omega)|) + IDFT(\log|H(\omega)|) \quad (7)$$

This process is better understood with the help of a block diagram (Figure 11). A lifter is used to separate the high quefrency (Excitation) from the low quefrency (Transfer Function). Figure 12 consists of the sounds 'eee' and 'aah' uttered by male and female speakers. We can see in the plot that the female speakers have higher peaks than the male speakers, which is due to higher pitch of female speakers. The initial 5ms consists of the transfer function and the later part is the excitation.

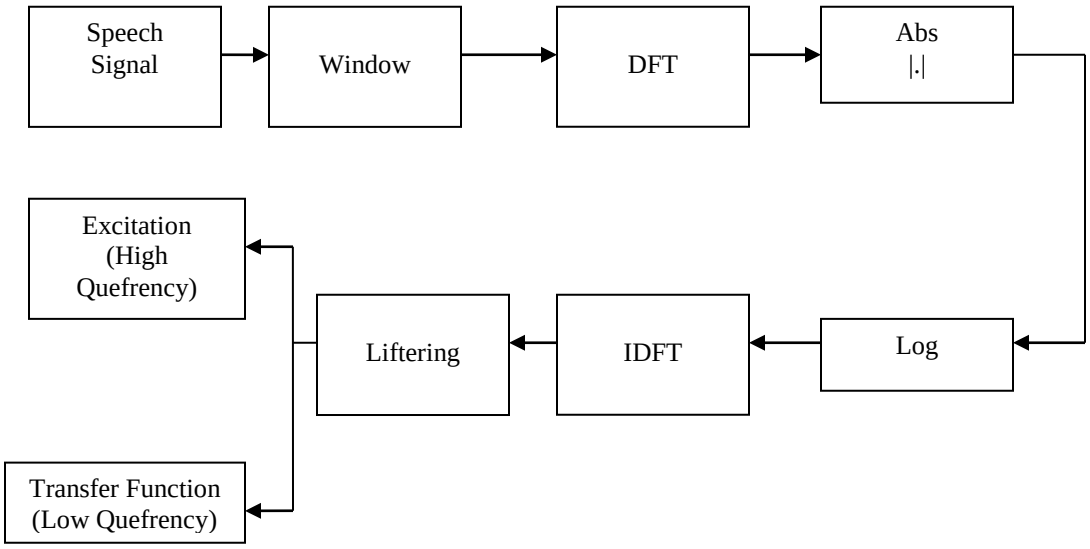


Figure 11. Block Diagram for Computing Cepstrum

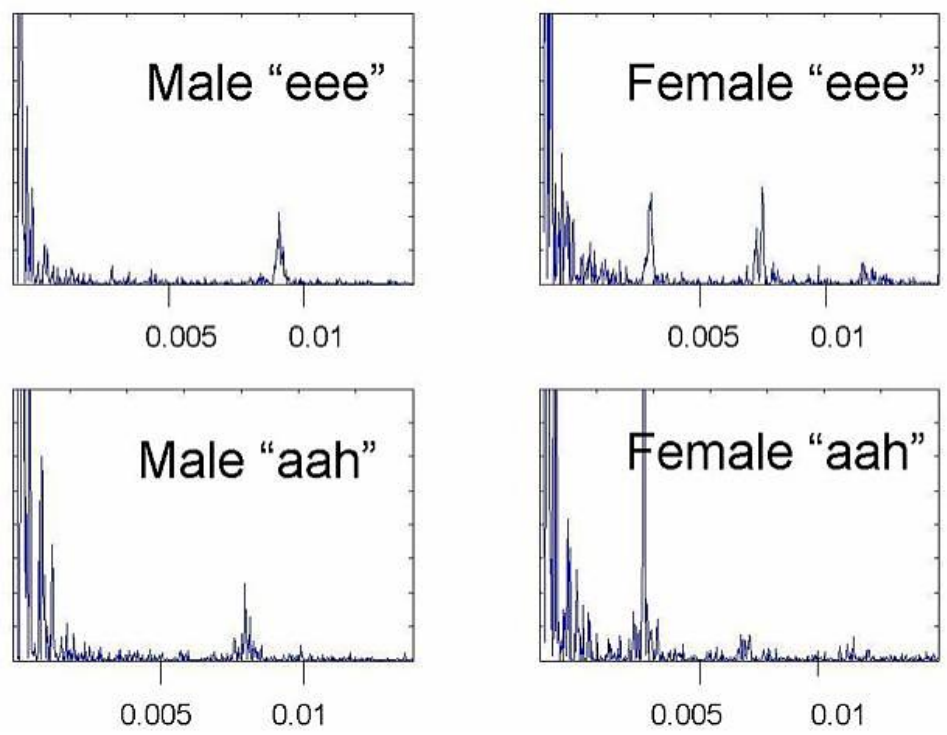


Figure 12. Cepstrum Plots

2.5. Feature Parameters

Now before we move on to the MFCC-GMM implementation, I would like to talk about the other more intuitive hardware implementation of other very well-known algorithms .The first thing that people did was try to distinguish speakers based on their pitch, which did not prove very successful as the pitch changes while one speaks and one can modulate their own voice to create a different pitch and hence one came across Formant estimation and deciding whether the speakers are same or not using this but still one found that although modeling using GMM gave a high enough accuracy the GMM-MFCC pair gave the highest accuracy on diverse conditions and also in presence of noise and also for a wide variety of speakers and hence my term paper will be focusing on this

particular area. So firstly we will be talking about formant basd speaker recognition and then we will be moving on to something call GMM and formant estimation fusion and lastly we will be talking about GMM and MFCC based speaker recognition.

2.5.1Pitch

Pitch is a perceptual attribute of sounds, defined as the frequency of a sine wave that is matched to the target sound in a psychophysical experiment Fundamental frequency (F0) is the corresponding physical term, defined only for periodic or nearly periodic sounds,F0 is the inverse of the period, in ambiguous situations, the period corresponding to perceived pitch is chosen. Usually pitch ≈ F0, both are measured in Hertz units.

Pitch estimation has been basically basically be done using autocorellation based algorithms, as mentioned in the papers(7),(8)..Autocorrelation function (ACF) based algorithms

Among the most frequently used F0 estimators. Usually the maximum value in ACF is taken as 1/F0 period Short-time ACF r(τ) for a discrete time domain signal x(n)

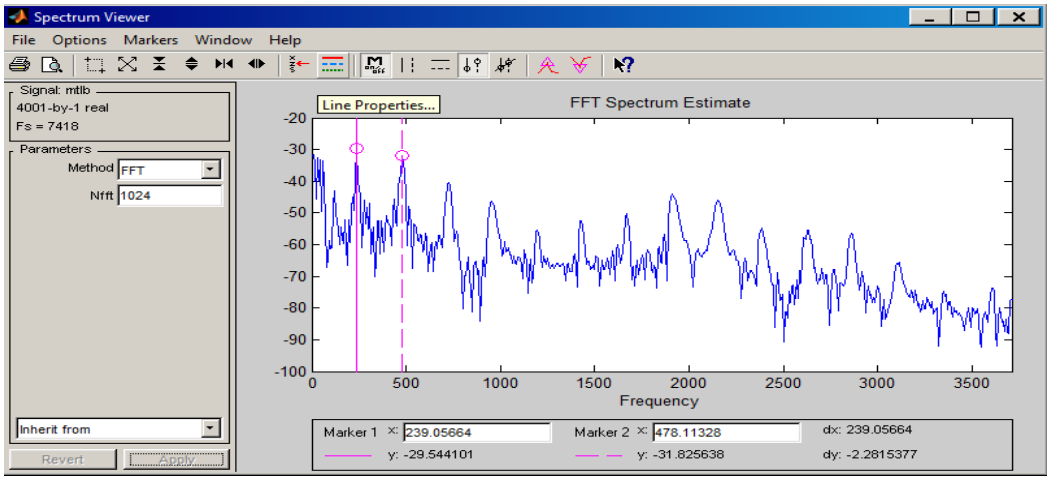


Figure 13. Pitch Plots

The figure above shows that the difference between the two peaks in the figure above of the fft of the speech signal is around 237 hz which gives us the fundamental frequency or the pitch of the speech signal.

The pitch detector mentioned below and first modeled as seen in the paper(8) can estimate the pitch at 10Khz sampling rate.. The centre clipping and peak clipping algorithm is very simple so if the speech signal crosses a certain part of the peak of the speech signal we don't allow that part of the speech signal and centre clipping is just thresholding at two directions and when we take the autocorrelation of these two functions we get something useful as we just get the pitch as peaks in the autocorrelation function hence using the sample delays we can estimate the pitch easily and the block diagram shown below does a wonderful job of estimating the pitch. Another thing that can be seen is that the has to be implemented is the buffering of the data, in the paper a small data buffer was used to segment the speech and calculate the pitch using that segment and it proved to be very accurate. The figure below illustrates the point I made. The other thing to be noted is that the autocorrelation function generated is periodic so we some however need to capture this periodicity so a comparator has been implemented and what this comparator does is compares the peaks of the autocorrelation and if the peaks don't vary by much of a fact then we can clearly say that the time interval between these two peaks multiplied by the sampling rate can give us the estimated pitch. The hardware implementation block diagram is shown in Fig(3) which tells us the basic building blocks that need to be used to create the pitch detection. The reason behind the fact that we get the periodicity is because of the fact that the speech signal has inherently inside of it a fundamental frequency. This was first noted quite a few years back by

Rabiner and Sodhi but then at that time no real effort was made to make this into an hardware circuit but later on rabiner was the force behind creating the hardware model which is still used now a days. The other thing to be noted is that the sampling rate plays an defining role, so for instance the change in sampling period due to some hardware circuit fault will cause some problems in the accuracy of the pitch detected and hence a lot of care has been taken in making the sampling frequency or the sapling frequency generator to be very stable as a wrong guess in the output can totally change of the identity of the person we were trying to determine. Now a days cepstrum is used to calculate the pitch which is sampling frequency independent.The figure below shows the use of different windowing period for the autocorrelation function for two window sizes one for a window size of 75 and the other for a window size of 36. Formants are thus a feature which can be utilized in generating and adding to some of the information needed for generating information which when modeled by the GMM is used for identifying speakers, the task at hand. It has been some work in the past that formants itself can be used to generate speaker recognition system. Also choosing more than 4 formants is not used generally, hence we have used 4 formants.

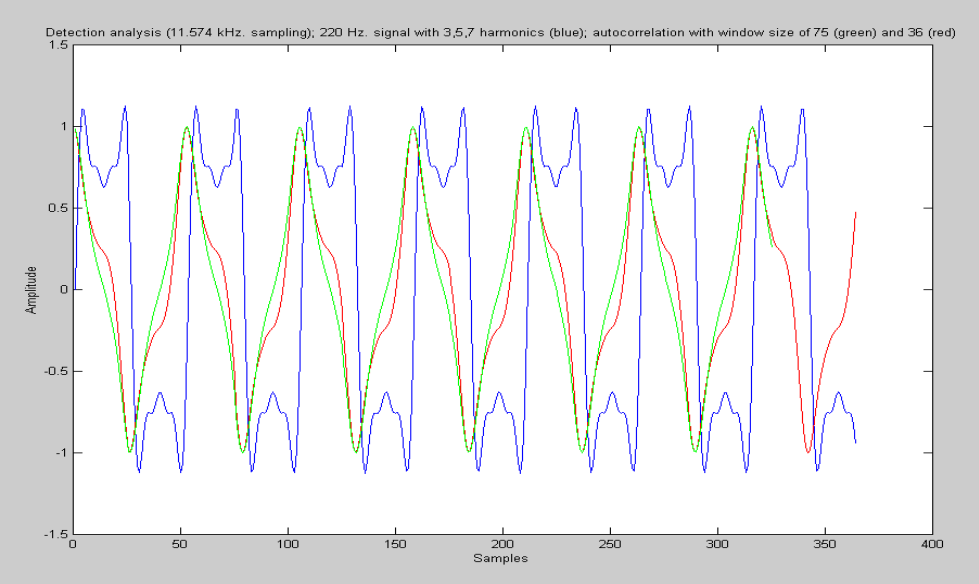


Figure 14. Formant Plots

The silence periods are also detected so as no to generate any false outputs. The LPF shown is a digital implementation of a 0-900 Mhz low pass filter.

2.5.2 Formants

Formant was formulated by Gunnar Fant (1960): 'There are spectral peaks of the sound spectrum which is defined here as |P(f)| are called *formants*'. This definition has been used in acoustics research and industry.

Formants are basically resonances occurring in the vocal tract. They generally occur every 1000 Hz and they basically is a concentration of acoustic energy at a particular frequency. There has been a lot of research done in this field to carry out speaker recognition using Formants and in 2008 the paper (5) tells us about speaker recognition using formants and they do it by estimating the formants and model them using GMM.

The figure below is a computer simulation I did using a person's speech and as can be seen the 3 peaks are the 3 formants. The 1st formant is at around 760 Hz and the next formant is at 2064 Hz and the third is around 2800 Hz. So the paper (5) basically talks about estimating the 1st three formants and then using Gmm classifying the data. Formants are also one major characteristic by which one can easily distinguish between the consonants and the vowels, hence there is something called a vowel triangle which is made up of the first three formants which and the location of the three formants can tell us if it is a vowel or a consonant. Thus formants in general are very powerful tools to generate the results for speaker as well as recognition of separate words.

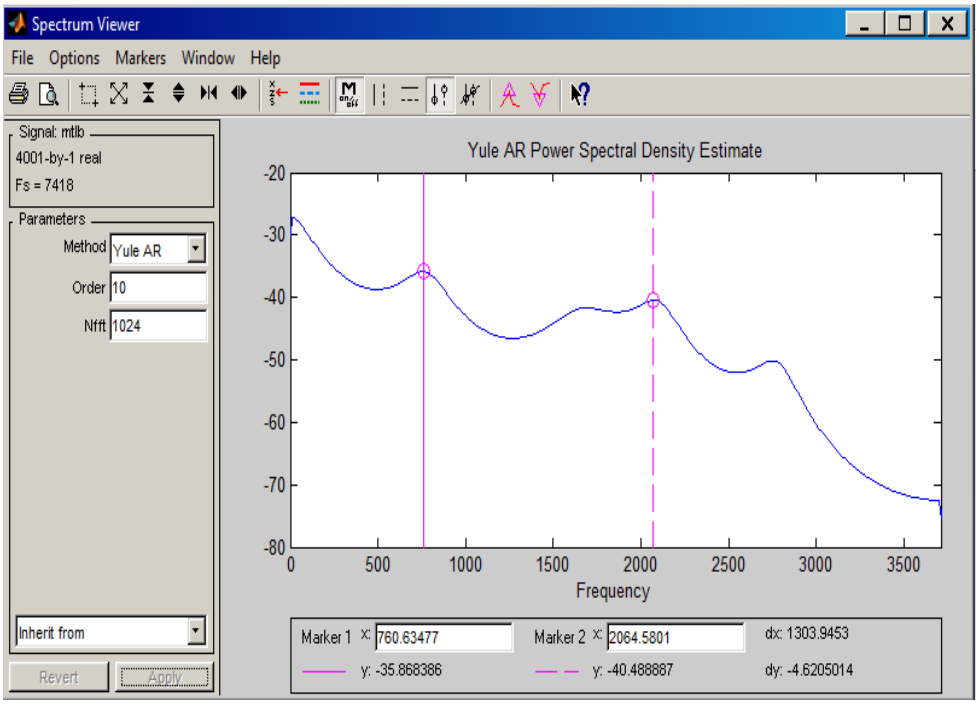


Figure 15. Formant Extraction

Estimation of the formants is generally a more difficult task than estimating fundamental frequency. The vocal tract system plays an role in producing the formants and hence

spectral shape of the vocal tract excitation plays an important role in the observed spectral envelope such that we cannot guarantee all vocal tract resonances will cause peaks in observed spectral envelope. The best way to model the speech formants is by modeling its as if it were modeled by some sort of source filter.

The premphesis stage just filters the speech and all the inverse filters are centred at particular frequencies and the output of the blocks generate the first 6 formants .

Then comes mixture modelling which can also be done using hardware and hence the above algorithm is highly implementable.

The thing to see here is that the hardware circuit assumes a 1000 Hz difference between the formants which leads to a certain amount of error in the output but that is not s such a problem as the relative difference is very low.

2.5.3MFCC

For speech/speaker recognition, the most commonly used acoustic features are mel-scale frequency cepstral coefficient (MFCC for short). MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. We shall explain the stey-by-step computation of MFCC in this section. Success has been due to representation of the speech amplitude spectrum into compact form.

The main steps for finding the Mfcc coefficients are taking the log magnitude spectrum of the windowed waveform which will then be smoothened out by triangular filters and then compute the DCT of the waveform to generate the mfcc coefficients.

Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition.

MFCC is a very robust parameter for modeling the speech as it can be modeled as MFCC kind of models the human auditory system and hence makes the reduction of the frame of a speech into the MFCC coefficients a very useful transform as now we have an even more accurate transform to deal with for the recognition of the speakers.

The other thing is that MFCC is one tool that produces very high level of accuracy when used as a parameter model for modeling the speech and hence for my study I have gven focus in this area.

The premphasis stage in the block diagram shown below basically does low pass filtering and then we will be windowing the data as it is known that speech is not an stationary process and one has to window the data to get some sort of accuracy.

The figure above gives the block diagram for computing the MFCC coefficients.

BASIC STEPS:

1)-Pre-emphasis

The speech signal s(n) is sent to a high-pass filter:

$$s_2(n) = s(n) - a*s(n-1) \quad (8)$$

where $s_2(n)$ is the output signal and the value of a is usually between 0.9 and 1.0. The z-transform of the filter is

$$H(z)=1-a*z^{-1} \quad (9)$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants. The next example demonstrates the effect of pre-emphasis.

2)-Frame Blocking

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is 320/16000 = 0.02 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second.

3)-Hamming Windowing

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame (to be detailed in the next step). If the signal in a frame is denoted by s(n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by the equation below.

$$w(n, a) = (1 - a) - a \cos(2pn/(N-1)), \quad 0 \leqq n \leqq N-1 \qquad (10)$$

4**)-Fast Fourier Transform** or **FFT**

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame. When we perform 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 in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies: Multiply each frame by a Hamming window to increase its continuity at the first and last points.

Take a frame of a variable size such that it always contains a integer multiple number of the fundamental periods of the speech signal. The second strategy encounters difficulty in practice since the identification of the fundamental period is not a trivial problem. Moreover, unvoiced sounds do not have a fundamental period at all. Consequently, we usually adopt the first strategy to mutiply the frame by a Hamming window before performing FFT.

5**)-Triangular Bandpass Filters**

We multiple the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpassfilter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation:

$$mel(f)=1125*\ln(1+f/700) \quad (11)$$

Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's subjective aural perception.

6**)-Discrete Cosine Transform** or **DCT**

In this step, we apply DCT on the 20 log energy E_k obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. The formula for DCT is shown next.

$$C_m=S_{k=1}^{N}\cos[m*(k-0.5)*p/N]*E_k,\ m=1,2,\ ...,\ L \quad (12)$$

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. Usually we set N=20 and L=13. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition. For better performance, we can add the log energy and perform delta operation.

DELTA MFCC equation is given below.

$$\Delta MFCC(K) = MFCC(K) - MFCC(K-1) \quad (13)$$

Here K is the coefficient number i.e K=1 to L.

So delta mfcc is kind of like taking the first derivative of the mel-cepstral coefficients.

So it tells us the rate of change of the cepstral coefficients, which will be useful in determining accents and not that useful in speaker verification systems. Still we will be demonstrating the delta mfcc in our proposed framework.

2.5.4 Shifted MFCC

Mel is a unit of measure of perceived pitch or frequency of the tone. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch in this linear manner. The Mel-scale is approximately linear below 1 KHz and logarithmic above. Here what we have done is shifted the frequency of windowing using triangular filters which is one of the steps for doing MFCC, instead of the normal 200-1500 Hz frequency windowing , we have done 1500-2500 Hz windowing, which has led to the addition of the accent features as mentioned above. The reason behind choosing this frequency range was to get the accent features, it has been found in various papers like[8] that the accent features are somewhat linked to this range of frequencies.

CHAPTER 3

ACCENT CLASSIFICATION SYSTEM

Before we proceed towards the features and modeling algorithms used in this system, a brief background and a research review on accent is presented in the next sections.

3.1 Accent Background

Foreign accent has been defined in [30]as the pattern of pronunciation features which characterize an individual's speech as belonging to a particular group. The term accenthas been described in [31] as – "The cumulative auditory effect of those features of pronunciation which identify where a person is from regionally and socially. In [32], Accent is described as the negative (or rather colorful) influence of the first language (L1) of a speaker to a second language, while Dialects of a given language are differences in speaking style of that language (which all belong to L1) because of geographical and ethnic differences.

There are several factors affecting the level of accent, some of the important ones are as follows:

1) Age at which speaker learns the second language.
2) Nationality of speaker's language instructor
3) Grammatical and phonological differences between the primary and secondary languages.

4) Amount of interaction the speaker has with native language speakers.

3.2 Review of Past Research

There has been considerable amount research of research conducted on the problem of accent modeling and classification. The following is a brief review on some of the papers published in the area of accent modeling and classification.

In[30], analysis of voice onset time, pitch slope, formant structure, average word duration, energy and cepstral coefficients was conducted. Continuous Gaussian Mixture GMMs were used to classify accents, using accent sensitive cepstral coefficients (ASCC), energy and their delta features. The frequencies in the range of 1500-2500 Hz were shown to be the most important for accent classification. A 93% classification rate was observed, using isolated words, with about 7-8 words for training. The Duke university dataset was used for evaluations. This dataset consists of neutral American English, German, Spanish, Chinese, Turkish, French, Italian, Hindi, Rumanian, Japanese, Persian and greek accents. The application was towards speech recognition and an error rate decrease of 67.3%, 73.3% and 72.3% from the original was observed for Chinese, Turkish and German accents respectively.In[33], fundamental frequency, energy in rms value, first (F1), second (F2), third formant frequencies (F3), and bandwidths of F1, F2 and F3; B1, B2 and B3 respectively were selected as accent features. The result shows the features in order of importance to accent classification to be: dd(E), d(E), E, d(F3), dd(F3), F3, B3, d(FO), FO, dd(FO), where E isenergy, d() are the first derivatives and dd() are the second derivatives. 3-state HMMs with single Gaussian densities were used for classification. A classification error rate of 14.52% was observed. Finally, they show an average 13.5% error rate reduction in speech recognition for 4 speakers by using

accent adapted pronunciation dictionary. The TIMIT and HKTIMIT corpuses were used as the database for evaluation. This paper was focused on Canto-English where their Cantonese is peppered with English words and their English has a particular local Cantonese accent. In [32]three different databases were used for evaluation: CU-Accent corpus – AE: American English, and accents of AE (CH: Chinese, IN: Indian, TU: Turkish), IviE Corpus: British Isles for dialects. CU-Accent Read – AE (CH: Chinese, IN: Indian, TU: Turkish) with same text as IviE corpus.A pitch and formant contour analysis is done for 3 different accent groups – AE, IN and CH (taken from CU-Accent Corpus) with 5 isolated words – 'catch', 'pump', 'target', 'communication', and 'look', uttered by 4 speakers from each accent group. Two phone based models were considered – MP-STM and PC-STM. The MFCCs were used as features to train and test STMs for each phoneme in case of MP-STM and phone class in case of PC-STM. Results show that better classification rate for MP-STM than PC-STM and also dialect classification was better than accent classification. The application was towards a spoken document retrieval system. In [34], LPC Delta cepstral features were used as features which were modeled by using 6 Gaussian mixture CHMMs. The classification procedure, employed gender classification followed by accent classification. A 65.48% accent identification rate was observed.The database used for evaluation was developed in the scope of the SUNSTAR European project. It consists of Danish, British, Spanish, Portuguese and Italian accents. In [35], a mandarin based speech corpus with 4 different accents was used as the native accent. A parallel gender and accent GMM was used to model, with 39 MFCCs consisting of 12 MFCCs and energy with first and second derivatives as features. Accent identification error rates of 11.7% and 15.5% were achieved for female and male

speakers respectively, using 4 test utterances and 32 component GMM. The application was speech recognition, but no experiment was conducted to test the performance on a speech recognition system.In[36], 13 MFCCs were used as features, with a hierarchical classification technique. The database was first classified according to gender, and 64-GMM was used for accent classification. They have used TI digits as the database and results show an average 7.1% error rate reduction relatively when compared to direct accent classification. The application was towards developing an IVR system using VoiceXML. In[37], 10 native and 12 non-native speakers were used as a dataset. Demographic data including speaker's age, percentage of time in a day when English used as communication and the number of years English was spoken were used as features, along with speech features: average pitch frequency and averaged first three formant frequencies. Even in this paper F2 and F3 distributions of native and non-native groups show high dissimilarity. Three neural network classification techniques namely competitive learning, counter propagation and back propagation were compared. Back propagation gave a detection rate of 100% for training data and 90.9% for testing data. The application was towards speech recognition, but an experiment has not been conducted in this regard.In[38], American and Indian accents have been extracted from the speech accent archive (SAA) dataset. Second and third formants were used as features and modeled with a GMM. The authors have manually identified accent markers and have extracted formants for specific sounds such as /r/, /l/ and /a/. They have achieved about 85% accent classification rate. In[39], speech corpus consisting of speakers from 24 different countries was used. The corpus focuses on French isolated words and expressions. Though this was not an application towards accent classification, this paper

showed that addition of phonological rules and adaptation of target vowel phonemes to native language vowel phonemes helps speech recognition rates. Also adaptation with respect to the most frequently used phonemes in the native languages resulted in an error rate reduction from 8.88% to 7.5% for foreign languages. An HMM was used to model the MFCCs of the data.In [40], the CU-Accent corpus, consisting of American English, Mandarin, Thai, and Turkish was used. 12 MFCCs along with energy were used as features and Stochastic Trajectory Model (STM) was used for classification. This classification employs speech recognition in front end, and was used to locate and extract phoneme boundaries. Results show that STM has classification rate of 41.93% when compared to CHMM and GMM which has 41.35% and 40.12% respectively. Also the paper lists the top five phonemes which could be used for accent classification. The application was towards speech and speaker recognition but know experiments have been conducted in this regard.

All the above accent classification systems were based on the assumption that the input text or phone sequence is known, but in our scenario where accent recognition needs to be applied to text-independent speaker recognition, a text-independent accent classification should be employed. In [40], text-independent accent classification effort has been made by using speech recognizer as front end followed by stochastic trajectory models (STM). However, this will increase the system complexityas well as introduce additional errors in the accent classification system due to accent variations.Our text-independent accent classification system comprises of a fusion of classification scores from continuous Gaussian hidden Markov models (CHMM) and Gaussian mixture models (GMM). Similar work has been done in the area of speaker recognition in[26],

where scores from two recognition systems were fused and one of the recognition algorithm was a Gaussian mixture model (GMM) and the other being a speaker adapted HMM instead of a CHMM.

3.3 Accent Classification Model

The AC model is as shown in Figure 17. Any unknown accent is classified by extracting the accent features from the sampled speech data and measuring the likelihood of the feature belonging to a particular known accent model. Any datasetwhere speech was manually labeled according to accents can be used as the reference accent database.

In this work, we have used a fusion of mel-frequency cepstral coefficients (MFCC), accent-sensitive cepstral coefficients (ASCC), delta ASCCs, energy, delta energy and delta-delta energy. Once these accent features have been extracted from the reference accent database (SAA dataset), two accent models are created with the help of GMM and CHMM. Any unknown speech is processed and accent features are extracted, then the log likelihood ofthose features against the different accent models are computed. The accent model with the highest likelihood score is selected as the final accent. In order to boost the classification rate the GMM and CHMM accent scores were fused. Due to the compensational effect [26] of the GMM and CHMM we have seen improvement in the performance.

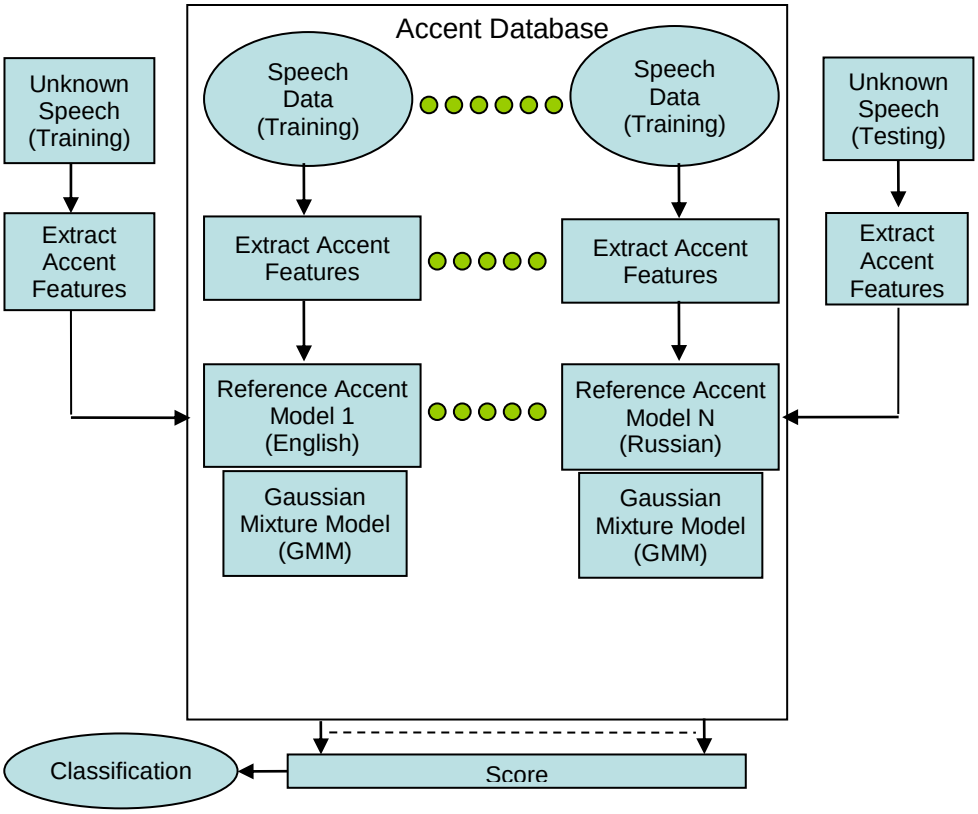


Figure 16. Block Diagram of Accent Classification (AC) System

3.4Accent Features

Researchers have used various accent features such as pitch, energy, intonation, MFCCs, formants, formant trajectories, etc., and some have fused several features to increase accuracy as well. In this paper, we have used a fusion of mel-frequency cepstral coefficients (MFCC), accent-sensitive cepstral coefficients (ASCC), delta ASCCs, energy, delta energy and delta-delta energy. MFCCs place critical bands which are linear up to 1000 Hz (Figure 18)and logarithmic for the rest. Hence it allows more selection filters on the lower 1000 Hz, whereas ASCCs [30], concentrate more on the second and third formants. i.e., around 2000 to 3000 Hz (Figure 19) which are more important features for detecting accent. Hence a combination of both MFCCs and ASCCs has been

used in this work which provided an increase in the accent classification performance when compared to ASCCs alone. Thus after these features are extracted, they are modeled using GMM and CHMM.

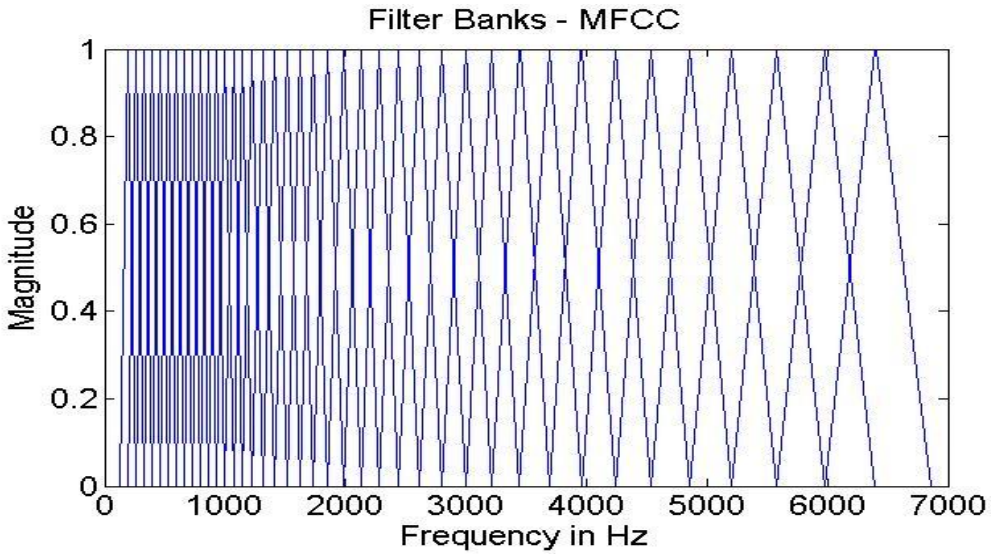
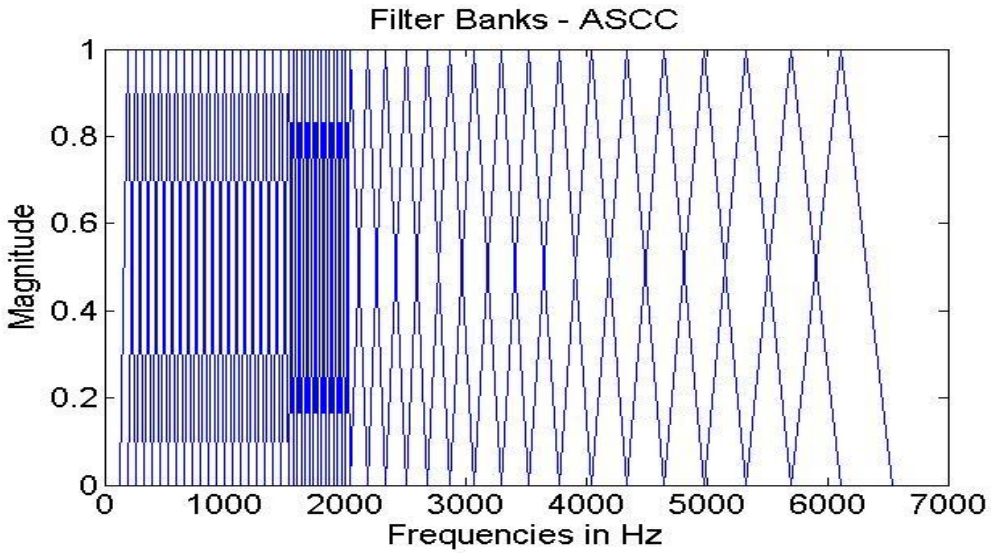


Figure 17. Mel Filter Bank



(b)

Figure 18. Accent Filter Bank

3.5 Accent Classifier Formulation

Gaussian mixture model (GMM) and has been used to achieve enhanced classification performance. GMM is explained here.

3.5.1 Gaussian Mixture Model (GMM)

A Gaussian mixture density is a weighted sum of M component densities which is given by,

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{14}$$

Where $\vec{x}$ is a D-dimensional vector, $b_i(\vec{x})$, $i = 1,\ldots,M$, are the component densities and p_i are the mixture weights. Each component density is given by,

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \sum_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \tag{15}$$

with mean vector modeling $\vec{\mu}_i$ and covariance matrix $\sum_i$. These parameters are represented by,

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\} \quad i = 1,...,M \tag{16}$$

These parameters are estimated iteratively using the Expectation-maximization (EM) algorithm. The EM algorithm estimates a new model $\bar{\lambda}$ from an initial model λ, so that the likelihood of the new model increases. On each estimation, the following formulae are used,

$$\bar{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \tag{17}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \tag{18}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)\vec{x}_t^{\,2}}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \bar{\mu}_i^{\,2} \tag{19}$$

where $\bar{\sigma}_i^2$, $\bar{\mu}_i$, and $\bar{p}_i$ are the updated covariance, mean and mixture weights. The a posteriori probability for class i is given by,

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{20}$$

For accent identification, each accent in a group of S accents, where S={1,2,….S}, is modeled by GMMs $\lambda_1, \lambda_2, ..., \lambda_S$. The final decision is made by computing the a posteriori probability for each test sequence (feature) against the GMM models of all accents, and selecting the accent which has the maximum probability or likelihood coefficients and then we model the data using multiple Gaussian curves and then what we do is to try to model the other datas which we want to compare using this reference data and find the distance using the GMM model[3][4][5].

CHAPTER 4

EXPERIMENTAL RESULTS

The HF system, accent classification system and the HFA system have been evaluated on various datasets; the results of these experiments are provided in this chapter. The HF speaker recognition system has been evaluated on YOHO[41] and the USF multi-modal biometric dataset[8]. For evaluating accent incorporation, i.e. accent classification system and HFA system, SAA system and the USF multi-modal biometric dataset were used. The YOHO dataset was not used for evaluating accent incorporation, as the dataset comprised of only north American accents.

4.1 TIDIGIT Dataset

The data set we used was of TIDIGIT for the purpose of designing and evaluating algorithms for speaker-independent recognition. There are 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 77 digit sequences. Each speaker group is partitioned into test and training subsets. A segment of TIDIGIT dataset was used. Total of 40 individual and different users in the dataset were used for enrollment and verification purposes The data set we used was of TIGIT, it is for the purpose of designing and evaluating algorithms for speaker-independent recognition of connected digit sequences. There are 326 speakers (111 men, 114 women, 50 boys and 51 girls)

each pronouncing 77 digit sequences. Each speaker group is partitioned into test and training subsets. A portion of TIDIGIT dataset was used. Total of 40 individual dataset was used for enrollment and verification purposes.

4.2.Results

We used MFCC, delta MFCC, the shifted MFCC, pitch and the first four formants as feature parameters and modeled using the GMM. Data is sampled at 16 kHz with 25 ms blocks and for each block, we extracted 13 MFCC, delta MFCC, and the shifted MFCC parameters. So in total including the pitch and the four formants, there are 44 parameters for each frame of the speech data. Also we used text dependent database to carry out our experiment. The results were averaged over 20 trials.

Figure 20 represents the recognition performance DET curve in terms of false rejection rate (FRR) versus false positive or acceptance rate (FAR). From Figure 20 it can be seen that at 10% false acceptance rate we achieved a false rejection rate of nearly zero but for a 5% false acceptance rate we achieved around 5% false rejection rate meaning that the algorithm incorrectly rejects only five out of hundred speakers.

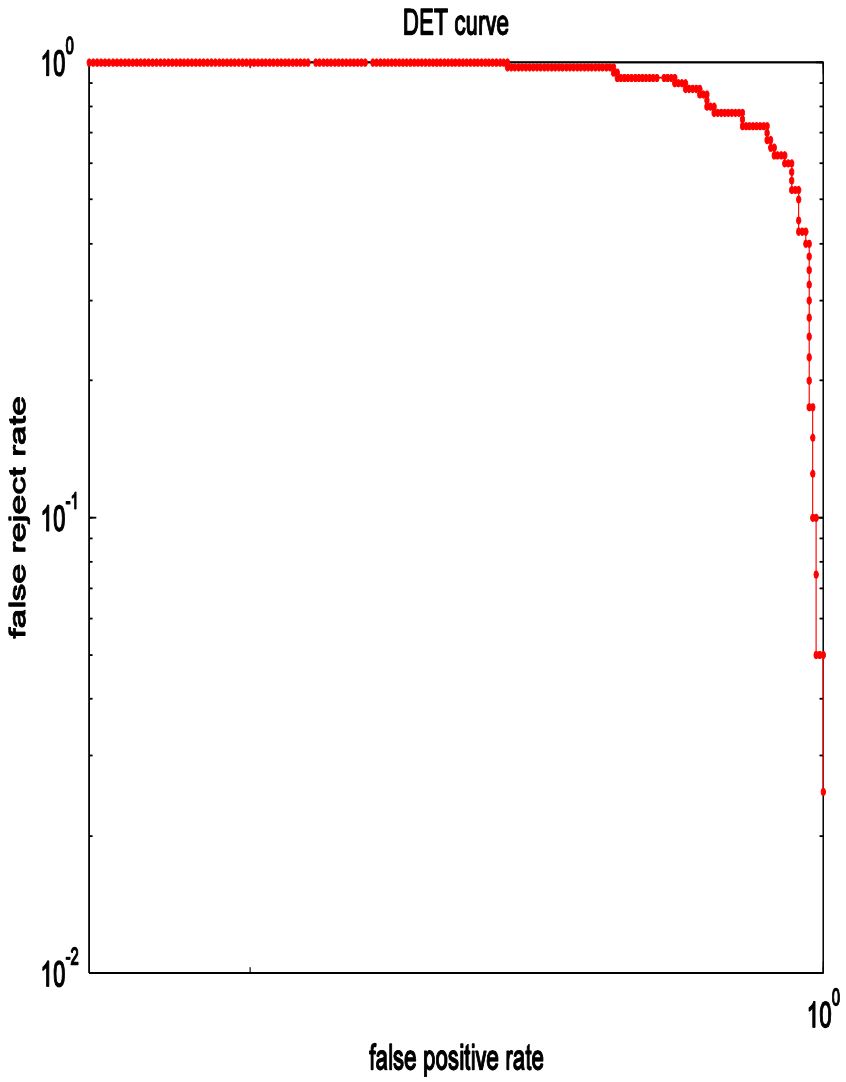


Figure 19. Recognition Performance DET Curve

Figure 19 represents the recognition performance ROC curve in terms of true positive or acceptance rate (TAR) versus false positive or acceptance rate (FAR). From Figure 19 it can be seen that at 10% FAR we achieved a 95% TAR (recognition accuracy) but for 5% FAR, the TAR is about 90%.

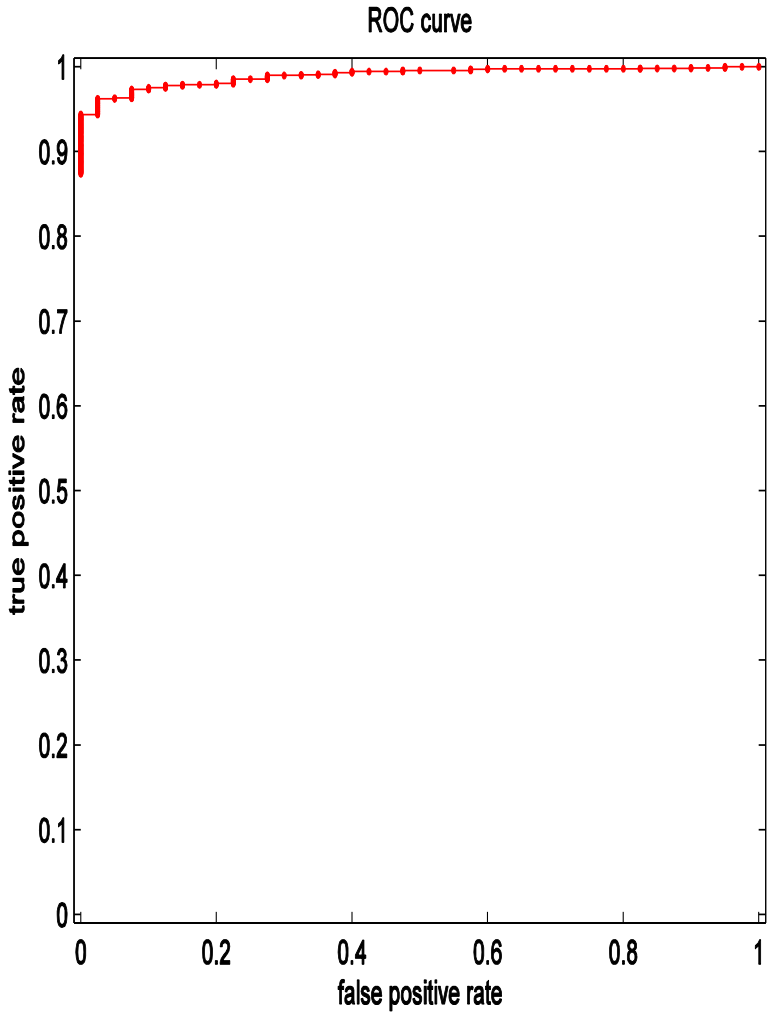


Fig 20. Recognition Performance ROC Curve

Based on our simulation results, we achieved approximately 10% improvement over the work shown in [1] and [2]. So we can conclude that the proposed algorithm improves the text dependent speaker recognition performance but at the cost of computational speed since we are modeling 44 speech parameters for each frame and hence it will definitely increases the computation time.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

So far we have seen that we have got a good level of accuracy also we will like to add more formant features and try to get a better level of accuracy as it is known that the more formant features are added the more accent features are added. A good biometric system needs to deliver a high performance at low FARs. By using a text-independent accent classification system with our HF system and score modifier algorithm, a significant enhancement has been achieved atlowFARs. In this thesis, speaker recognition using arithmetic harmonic sphericity (AHS) and hidden Markov model (GMM) has been studied. Mel-frequency cepstral coefficients (MFCC) have been used as speaker features.

A novel method to recognize/identify speakers was developed by including a new set of features, the shifted MFCC which allowed inclusion of accent information in the recognition algorithm. The algorithm was evaluated using TIDIGIT dataset and the results showed on the average 10% improvement over the performance of our previous work [2]. Our future work will be to evaluate the proposed feature set for a even more noisy database and and observe the impact on its performance. We need to study the effect of adding more formants to see if that will provide better accent discrimination. Also we need to investigate ways to improve the computational time complexity of the

proposed algorithm. For the first time a text-independent accent classification (AC) system has been developed without the usage of an automatic speech recognizer. MFCCs, accent sensitive cepstral coefficients (ASCCs) and energy have been used as accent features. MFCCs emphasize the first formant frequency, whereas ASCCs emphasize second and third formants. By combining MFCCs and ASCCs along with energy increases accent classification rate .Then, the speaker recognition system was combined with accent classification system to enhance the true acceptance rate (TAR) at lower false acceptance rates (FAR). The AC system produces accent class information and the accent score assigned to each speaker.

5.2 Recommendations

Complete automation of the accent classification system and the score modifier, would be useful, so that no tuning needs to be done for different datasets. Higher level features other than mel-frequency cepstral coefficients (MFCC), accent-sensitive cepstral coefficients (ASCC), delta ASCCs, energy, delta energy and delta delta energy needs to be integrated into the system, so that an accent classification rate can be improved which would enhance the Speaker Recognition system performance in turn. The Speaker Recognition system needs to be evaluated on a variety of larger datasets, so that more inferences can be drawn from the results and enhancements to the Shifted MFCC can be made. Also different fusion techniques at the modeling level such as SVM Vs. GMM, HMM Vs. SVM needs to be studied, and evaluated on a variety of datasets to better understand the effect of different fusions, so that a common technique can be formulated

to find the optimal fusion weights. Finally, as we know from the results that accent incorporation enhances speaker recognition, studies have to be conducted on several other factors such as gender classification systems. The process of identifying human through speech is a complex one and our own human recognition system is an excellent instrument to understand this process. The human recognition system extracts several other features from a single speech signal, due to which it achieves high accuracy. The goal of a speech researcher should be to identify such missing pieces of information, in a hope to match the human recognition system someday.

REFERENCES

[1] "Homeland Security Advisory System," *Journal*, [online], (Date), Available http://www.dhs.gov/xinfoshare/programs/Copy_of_press_release_0046.shtm.

[2] F. Bimbot and L. Mathan, "Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure," *Third European Conference on Speech Communication and Technology*, 1993.

[3] L. E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554-1563, 1966.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Proc.*, vol. 3, no. 1, pp. 72-83, January 1995.

[5] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. 32, no. 2, pp. 307-309, March 1986.

[6] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *Journal of Acoustical Society of America*, vol. 35, pp. 354-358, 1963.

[7] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical techniques for talker identification," *Journal of Acoustical Society of America*, vol. 50, pp. 1427-1454, 1971.

[8] B. S. Atal, "Text-independent speaker recognition," *Journal of Acoustical Society of America*, vol. 52, 1972.

[9] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines," *Proc. ICASSP*, pp. 524-527, 1989.

[10] H. Gish and M. Schmidt, "Text-independent speaker identification," *Signal Processing Magazine, IEEE*, vol. 11, no. 4, pp. 18-32, 1994.

[11] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods usingVQ-distortion and discrete/continuous HMM's," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 3, pp. 456-459, 1994.

[12] T. J. Hazen, D. A. Jones, A. Park, L. C. Kukolich, and D. A. Reynolds, "Integration of Speaker Recognition into Conversational Spoken Dialogue Systems," *Proc. Eurospeech*, pp. 1961-1964, 2003.

[13] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM," *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, 2004.

[14] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, 2004.

[15] M. M. Tanabian, P. Tierney, and B. Z. Azami, "Automatic speaker recognition with formant trajectory tracking using CART and neural networks," *Electrical and Computer Engineering, 2005. Canadian Conference on*, pp. 1225-1228, 2005.

[16] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete – Time Processing of Speech Signals*, NJ: IEEE Press, 2000.

[17] S. Gray and J. H. L. Hansen, "An Integrated Approach to the Detection and Classification of Accents/Dialects for a Spoken Document Retrieval System," *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pp. 72-77, 2005.

[18] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, vol. 1, 1999.

[19] C. Teixeira, I. Trancoso, A. Serralheiro, and L. Inesc, "Accent identification," *SpokenLanguage, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, 1996.

[20] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 343-346, 2001.

[21] X. Lin and S. Simske, "Phoneme-less hierarchical accent classification," *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, vol. 2, 2004.

[22] M. V. Chan, X. Feng, J. A. Heinen, and R. J. Niederjohn, “Classification of speech accents with neural networks,” *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 7.

[23] S. Deshpande, S. Chikkerur, and V. Govindaraju, “Accent classification in speech,” *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, pp. 139-143, 2005.

[24] K. Bartkova and D. Jouvet, “Using Multilingual Units for Improved Modeling of Pronunciation Variants,” *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5, pp. 1037-1040, 2006.

[25] P. Angkititrakul and J. H. L. Hansen, “Advances in Phone-Based Modeling for Automatic Accent Classification,” *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, vol. 14, no. 2, pp. 634-646, 2006.

[26] J. P. Campbell Jr, “Testing with the YOHO CD-ROM voice verification corpus,” *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, 1995.

[27] *Speech Accent Archive*, George Mason University, [online] Available http://accent.gmu.edu.

[28] Srikanth,"Voice recognition system based on intra-modal fusion and accent classification" Masters thesis(2007),USF