GAMMATONE AND MFCC FEATURES IN SPEAKER RECOGNITION

by

Wilson Burgos

Bachelor of Science

Computer Engineering

A thesis submitted to
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Computer Engineering

Melbourne, Florida
November 2014

We the undersigned committee
here by approve the attached thesis

GAMMATONE AND MFCC FEATURES IN SPEAKER RECOGNITION

by
Wilson Burgos

Veton Z. Këpuska, Ph.D.
Associate Professor
Electrical and Computer Engineering
Committee Chair

Samuel P. Kozaitis, Ph.D.
Professor
Department Head
Electrical and Computer Engineering

Marius C. Silaghi, Ph.D.
Assistant Professor
Computer Science

Abstract

Title: GAMMATONE AND MFCC FEATURES IN SPEAKER RECOGNITION
Author: Wilson Burgos
Committee Chair: Veton Z. Këpuska, Ph.D.

The feature analysis component of an Automated Speaker Recognition (ASR) system plays a crucial role in the overall performance of the system. There are many feature extraction techniques available, but ultimately we want to maximize the performance of these systems. From this point of view, the algorithms developed to compute feature components are analyzed. Current state-of-the-art ASR systems perform quite well in a controlled environment where the speech signal is noise free. The objective of this thesis investigates the results that can be obtained when you combine Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) as feature components for the front-end processing of an ASR.

The MFCC and GFCC feature components combined are suggested to improve the reliability of a speaker recognition system. The MFCC are typically the *"de facto"* standard for speaker recognition systems because of their high accuracy and low complexity; however they are not very robust at the presence of additive noise. The GFCC features in recent studies have shown very good robustness against noise and acoustic change. The main idea is to integrate MFCC & GFCC features to improve the overall ASR system performance in low signal to noise ratio (SNR) conditions.

The experiment are conducted on the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and the English Language Speech Database for Speaker Recognition (ELSDR) databases, were the test utterances are mixed with noises at various SNR levels to simulate the channel change. The results provide an empirical comparison of the MFCC-GFCC combined features and the individual counterparts.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

Acknowledgements

I would like to thank Dr. Veton Këpuska for helping me and steering my research in the area of Speaker Recognition. The opportunity to work with him, not only during my research but as his teaching assistant has helped me develop into a better overall engineer. Many thanks to Dave Straka for proofreading my work, it was of great help.

Alba Luz, my wife, thank you for all your love,
patience and support throughout all these years.
It has been an adventure and I would do it again if I had to.
Ian and Jared, my children, thank you.
You guys are awesome.
My parents Pablo and Carmen,
I have succeeded because of you.
And most importantly GOD, thank you for saving me twice

1 INTRODUCTION TO SPEAKER RECOGNITION

The human auditory system is very unique and becomes functional at around 25 weeks' gestation [1], even before birth our brain is being set up to learn a language. It's fascinating to know that so early on we start learning to recognize the things we hear. One of the very first things we recognize are the voices of our parents; especially our mother's voice [2]. This could be explained by the fact that we communicate with speech, which is one of the most natural methods of communicating [3]. However, this naturally learned process of recognizing different speakers presents many challenges when applied to Automatic Speaker Recognition Systems (ASRS).

Generally Speaker Recognition is often confused with Speech Recognition; they are related because both utilize speech signals, but they are significantly different. Speaker Recognition tries to figure out **who** was speaking whereas in Speech Recognition the goal is to find out **what** was spoken. Thus the goal of an ASRS in this context is to correctly identify or verify the speaker, effectively a biometric authentication. A Biometric Authentication is an automated method of verifying or recognizing the identity of a person based on a physiological or behavioral characteristic [4]. Chapter 2 provides more insights into the difference between speaker identification and speaker verification. However, the focus of our study will be in Speaker Identification.

The idea sounds interesting but there are intricate problems to accurately identifying an individual. Some of these issues will be discussed in Chapter 3. Nonetheless, Speaker Recognition is a popular choice for remote authentication due to the availability of devices to collect speech samples (phones & computers) [5] .

1.1 History of ASR Systems

1.1.1 1960s and 1970s

The development of Automated Speaker Recognition Systems has been relatively steady over the last five decades; always a step behind advancements in speech recognition. One of the first attempts to develop a system to recognize speakers automatically happened during the 1960's era. During that time researchers at the famous Bell Telephone Laboratories (Bell Labs), one of the premier facilities for research, were among the first to create a similarity measurement using an array of filter banks. Researchers were able to cross correlate the two signal spectrograms to determine how similar they were [6]. Subsequent studies improved this technique by analyzing the variances of a subset of features instead of all of them [7].

Meanwhile researchers at the Systems Development Division of similarly recognized International Business Machines Corporation (IBM) tried to perform speaker discriminations using an adaptive system. They were able to achieve significant improvements by applying linear discrimination analysis. In fact, they achieved over 90% accuracy distinguishing a known speaker from impostors [8].

Back at Bell Labs, George Doddington decided to use a different technique: *formant analysis*. He followed advancements in speech recognition that were based on acoustic phonetics which tried to model the vocal tract using a mathematical filter [9]. The experiment used these formant frequencies, pitch and speech energy to aid in verification decreasing the error rates by four times using existing filter bank techniques [10].

In 1976, Texas Instruments (TI) built the first fully automated speaker verification system. It was tested by the U.S. Air Force and the MITRE Corporation [11]. The verification was based on a pseudo-random 4 word phrase using digital filter banks for spectral analysis. Several millions of tests were completed using a sample of 200 speakers over a period of 6 years [12].

1.1.2 1980s and 1990s

In the beginning of the decade, Bell Labs created new experimental systems designed to work over the telephone lines. A researcher proposed using frame based features combining the cepstral coefficients and their 1st and 2nd polynomial to increase robustness [13] . These features later became a standard for both speech and speaker recognition.

In the mid-1980s a Speech Group was developed to promote and study new speech processing techniques by the National Institute of Standards and Technology (NIST) [5]. As new techniques for speech recognition were discovered, different alternatives to template matching were applied, and Hidden-Markov Model (HMM) based text dependent methods came to fruition. These new systems used speaker models derived from multi-word sentences combining them according to a specific sentence level grammar [14].

The features vectors that describe a unique speaker can be effectively compressed, utilizing clustering techniques such as Vector Quantization (VQ), into VQ codebooks. The VQ codebook is a small set of numbers that represent the feature vector dataset [15].

In the 1990s the focus of research was concentrated on increasing robustness and presenting Speaker Recognition as a viable biometric technology. The continuous HMM method was found to be very robust when enough training data was available, comparable to VQ-based methods [16]. A different technique was tried as well, combining the spectral envelope (Ceptral features) along with the fundamental frequencies. The method used two separate VQ-codebooks, for each speaker, one for voiced utterances and the other for unvoiced utterances; leading to increased recognition accuracy [17].

The input used for ASR is always contaminated by noise, which in turn introduces distortions and loss of information [18]. The HMM was adapted to handle noisy speech using the Parallel Model Combination (PMC) method, effectively adding an HMM model to the background noise. The speaker decision was made using the highest likelihood value for the model [19].

The NIST Speech Group, funded by the National Security Agency (NSA) has hosted biennially the Speaker Recognition Evaluations (SRE) since 1996. Their goal is to drive the technology forward while finding the most promising approaches [20].

1.1.3 2000s

One of the problems still faced in Speaker Recognition is dealing with the intra-speaker variations. Such variations can arise for multiple reasons such as: recording conditions, environment or mood, etc… One cannot assume a speaker can repeat an utterance in the same manner from trial to trial, that's where score normalization comes into place. Newer techniques have been proposed to normalize using the Z-Score, subtracting the mean and dividing by the standard deviation of the imposter score distribution [21].

Different kinds of modeling techniques have also been investigated utilizing Gaussian Mixture Models (GMM) and Support Vector Machine (SVM) Classifiers. The GMM mean super vector SVM system represents acoustic observations as a series of GMM vectors with discriminative SVM classification [22]. Other methods try to deal with the unsupervised adaption problem, which tries to update the client model using the test data. The adaptation includes a weighting scheme of the test data, based on the *a posteriori* probability that it belongs to a particular client model [23].

Recently there has been some interest to incorporate audio and visual components for ASR Systems, in which a combination of speech and image speech information is used. The typical visual information captured is the lip movement. This combination helps improve the reliability of the system; for instance the background noise does not affect the movement of the lips, but it has a detrimental effect on the voice quality. Conversely, lightning conditions do not have any effect on the voice performance but it does affect the quality of the lip recognition engine [24].

1.2 Summary of the Technology progress in ASR

Figure 1-1 summarizes some of the technological progress in the last 50 years. Research in the field of speaker recognition continues worldwide; advances in hardware and signal processing continues to spur innovation in this field.

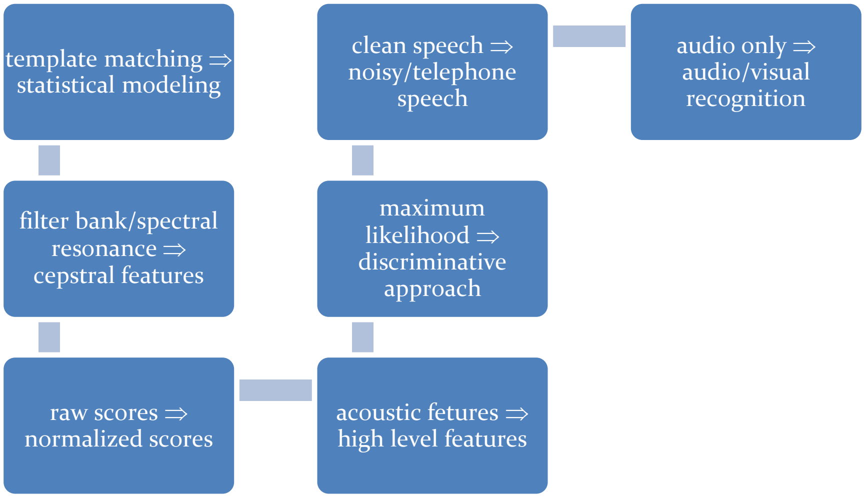

template matching ⇒ statistical modeling
clean speech ⇒ noisy/telephone speech
audio only ⇒ audio/visual recognition
filter bank/spectral resonance ⇒ cepstral features
maximum likelihood ⇒ discriminative approach
raw scores ⇒ normalized scores
acoustic fetures ⇒ high level features

Figure 1-1 – Summary of Technology progress in ASR

There have been many other techniques not included in the diagram, many of these improvements were geared toward increasing robustness and as such have spanned both fields, speech and speaker recognition [16].

Despite the numerous achievements that have been made over the years, we are just scratching the surface of the capabilities of new ASR systems as a viable biometric authentication system.

2 CHARACTERISTICS OF ASR SYSTEMS

An ASR system primarily tries to model the vocal characteristics of a person by using either a mathematical or a statistical model. In other words, a pattern recognition problem in which the input data is labeled for later classification. Once the model is established and associated with an individual, new utterances of speech may be analyzed to determine the likelihood of them being generated by the model in question or not. This is the underlying methodology under which all of the ASR systems operate and it typically it has two phases: Enrollment (front-end) and Verification (back-end). During the enrollment phase, the speaker's voice is analyzed, then a number of features are extracted to create a voice model of the speaker. The verification phase uses the voice model previously created to compare against a speech utterance.

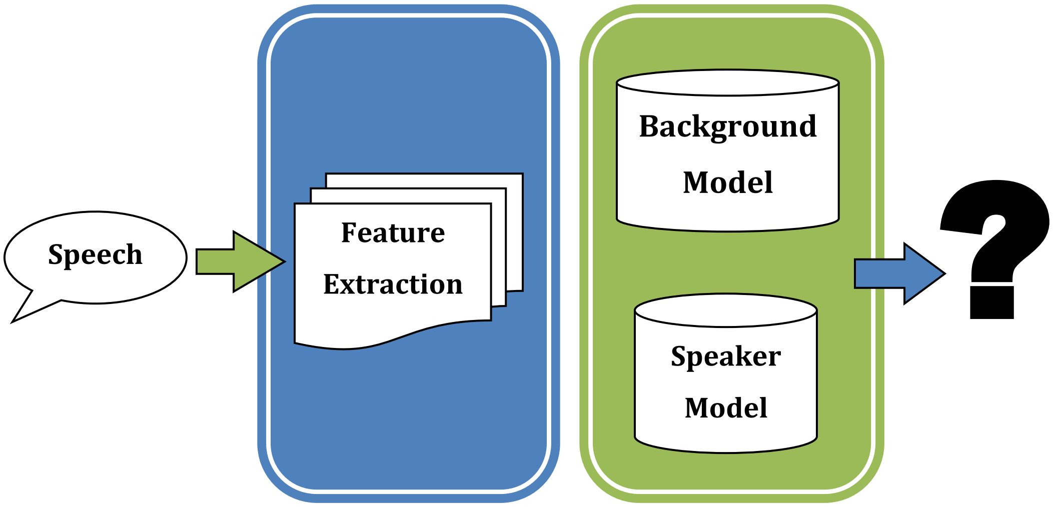


Figure 2-1 Block diagram of ASR

In the movie *Mission Impossible III*, Tom Cruise tries to fool the most common speech recognition engine, the human perception. He tries to do so by putting on a mask of Philip Seymour Hoffman, typical of *Mission Impossible* movies, but he forces the person to read something (enrollment phase) and uploads the audio to a remote computer which builds a model of the person's voice [25]. The parameters of the model are then sent to a device on Tom Cruise's neck, over his trachea, adaptively modifying his vocal characteristics to mimic Hoffman's voice, thus fooling others (verification phase).

More strictly speaking speaker recognition is the process of identifying the person who is speaking by the characteristics of their voices, *voice biometrics*. The key component of an ASR is precisely this back-end.

Generally speaker recognition can be classified into two areas: identification and verification.

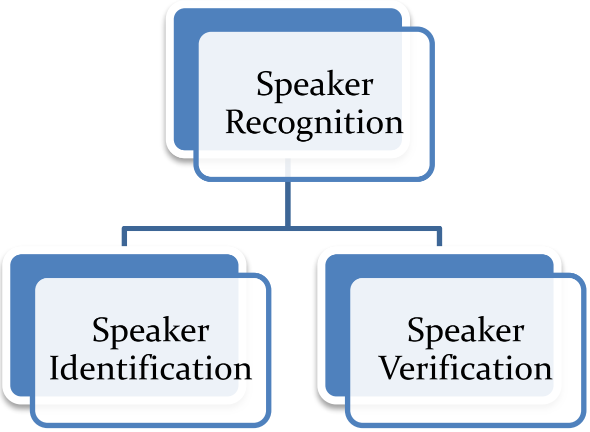


Figure 2-2 – Areas of Classification in a Speaker Recognition System

2.1 Speaker Identification

The purpose of speaker identification, in general, is to determine who is talking out of a group of known speakers [26]. The process of identification is similar to what a police officer does when comparing a sketch of a suspect against photos of criminals to find the closest match. There are two different types of speaker identification processes: *closed-set identification* and *open-set identification*. The closed-set identification task is simpler than the other. In closed-set identification the system has to perform a 1:N classification with the assumption that the unknown voice came from the set of known speakers [27]. The system will always come up with an answer, that being determined by the closest match. It is worthwhile to notice that in closed-set identification there is no rejection scheme. There will always be a speaker in the model closest to the unknown speaker [25].

In practical terms we could devise a test where the “unknown speaker” is an 8-year old female child, and all of the speakers in the database are adult males. The child will still match against one of the speakers in the database. On the other hand the open-set identification task can be seen as a combination of the closed-set identification and speaker verification. The first stage of the open-set identification is similar to that of closed-set identification, but the second stage uses speaker verification to decide whether the utterance was actually produced by the most likely speaker or an impostor [28]. The verification process can be defined as a special-case of the open-set identification in which there is only a single speaker in the list.

2.2 Speaker Verification

One can easily recollect a time in an airport, presenting our passport to a border control agent who in turn compares your face to the picture in the document; this is a verification process. Speaker Verification, also known as speaker authentication, is the task that determines if the speaker is who he claims to be. Sometimes this task is known as speaker authentication or voice verification. During the verification process it is generally assumed that imposters, those claiming to be a valid user, are not known to the system [26].

The verification task can be considered a binary decision problem because its output is either true or false. In the open-set *verification* task, the system distinguishes a voice that is known to the system, the claimers identify, from a potentially large group of voices unknown to the system.

These days in most ASR systems, you can find the verification task as the centerpiece, because of its commercial viability in different security applications, such as telephone access control for banking services. This task could happen in two variants; either *text-dependent* or *text-independent*.

2.2.1 Text-dependent vs Text-independent

Text-dependent refers to a system that requires the speaker to use the same text during training and test phases. Text-independent has no constraint on the speech content. Comparing them, a text-independent system is more convenient and commercially attractive, given the fact that the user can speak freely to the system. However, the main tradeoff is that in order to achieve better performance, longer training of test utterances are required [29].

2.3 Universal Background Model (UBM)

The task to detect a speaker could be defined as two hypothesis tests. The first test is the one in which the speech signal Z does come from the hypothesized speaker and the second one where it does not come from the hypothesized speaker. This can be defined as follows

H_1: Z does come from the hypothesized speaker S

and

H_2: Z does **not** come from the hypothesized speaker S.

The *likelihood* of the hypothesis H_i given the speech signal can be defined as the probability density function $p(Z \mid H_i)$. Then we can use a likelihood ratio test given by the two hypotheses to determine the decision. Mathematically each hypothesis is represented by a model for each speaker S in the feature space. The model for H_2 according to Reynolds [30] is represented by

$$p(X \mid \lambda_{\overline{hyp}}) = f\big(p(X \mid \lambda_1), \dots , p(X \mid \lambda_N)\big), \tag{2-1}$$

where *f* is the average or maximum function of the likelihood values from the background speaker set. This approach is called the *universal background model* (UBM), where we train a single model from several speakers representative of the total population of the expected speakers.

2.3.1 Gaussian Mixture Model

The likelihood function, $p(X \mid \lambda)$ selected for calculating the likelihood ratio of the model is very important. For text-independent speaker recognition the most successful one has been the Gaussian mixture models (GMM) [30]. A GMM could be thought of as a Gaussian distribution describing a one-dimensional random variable *X*. The variable *X* is defined as a vector described by the mean and variance. The mixture density for a feature vector, *X* can be defined as

$$\boldsymbol{p(X \mid \lambda) = \sum_{i=1}^{M} w_i p_i(X)} \tag{2-2}$$

This mixture density is a weighted linear combination of unimodal Gaussian densities, $p_i(X)$

$$p_i(X) = \frac{1}{(2\pi)^{D/2} \ \ |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{2-3}$$

The Figure 2-3 shows an example of a one-dimensional probability distribution that is more effectively modeled using a GMM with 3 mixtures.

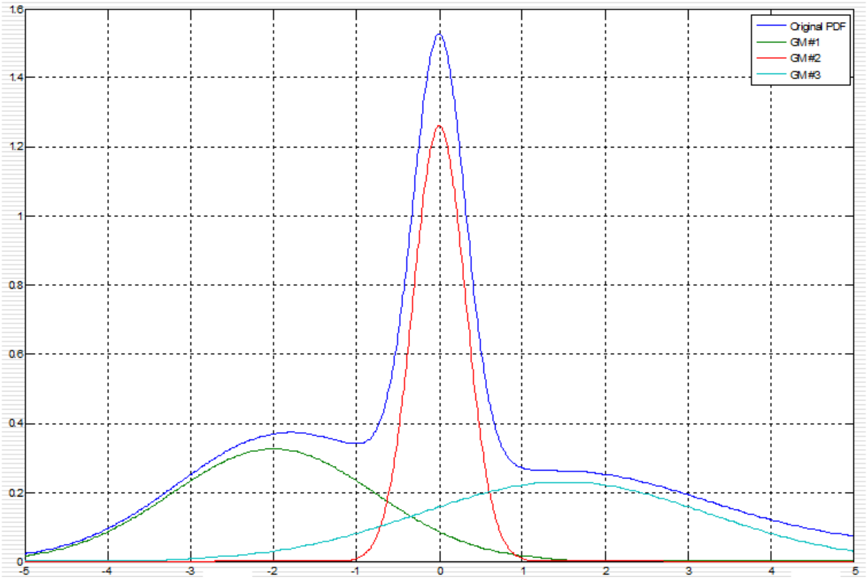


Figure 2-3 A Complex distribution using a 3 mixture GMM [31]

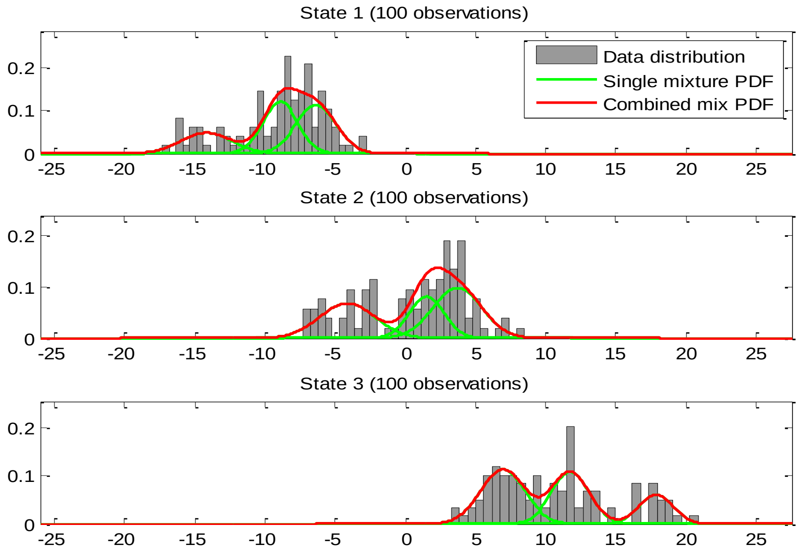


Figure 2-4 Training sequence with 3 Gaussian mixtures **[31]**

2.3.2 Expectation-Maximization

The UBM is trained using the Expected-Maximization (EM) algorithm. The EM algorithm refines the parameters of the GMM iteratively to increase the likelihood of the estimated model for the feature vectors being observed. According to Reynolds [30], generally five iterations are sufficient for convergence, for iterations *i* and *i+1*, $p(X\,|\lambda_{i+1}) > p(X\,|\lambda_i)$.

The log-likelihood of a model λ for a sequence of vectors, $X = \{x_i, \dots\, x_T\}$, is computed using ,

$$\log p(X\,|\lambda) = \sum_{t=1}^{T} \log p(x_t\,|\,\lambda) \qquad \textbf{(2-4)}$$

where $p(x_t\,|\,\lambda)$ is computed using equation (2-2), and the output is averaged to normalize any duration effects from the log-likelihood.

The simplest approach to train the UBM is to select samples from all the data, however the samples have to be selected carefully such that they are well balanced over the subpopulations of the total data.

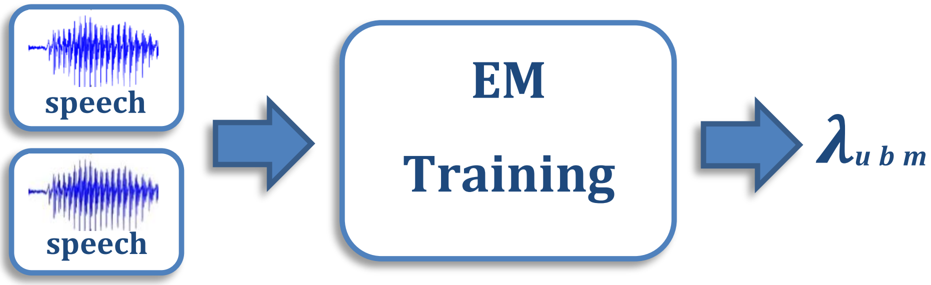


Figure 2-5 Data from subpopulations are pooled before training.

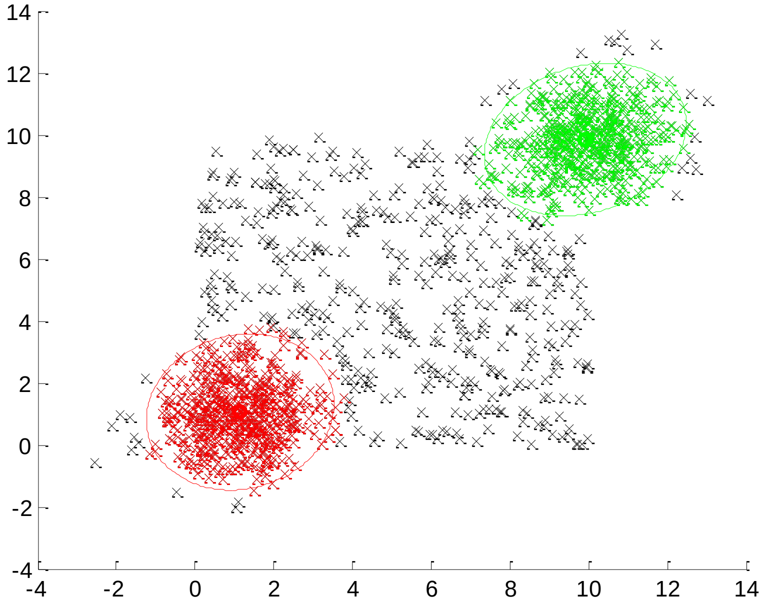


Figure 2-6 EM Algorithm convergence of clusters

2.3.3 Speaker Model Adaption

The speaker-specific model is adapted from the UBM using the maximum a posteriori (MAP) estimation. The adaptation increases the performance and provides a tighter coupling between the two models. The process follows a similar approach to the EM algorithm but instead the new statistics are combined with the statistics from the UBM. According to [30] the alignment of the training vectors to the UBM (Figure 2-7a) can be computed as follows

$$Pr(i \mid x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \qquad (2\text{-}5)$$

To compute the statistics for the weight, mean and variance parameters $Pr(i \mid x_t)$ and x_t is used. Once enough training statistical data has been gathered, the old UBM statistics are updated (Figure 2-7b). Once created the adapted mixture parameters use adaption coefficients to control the balance between old and new estimates. The parameters and equations for the general MAP estimation and its constraints are described in [30].

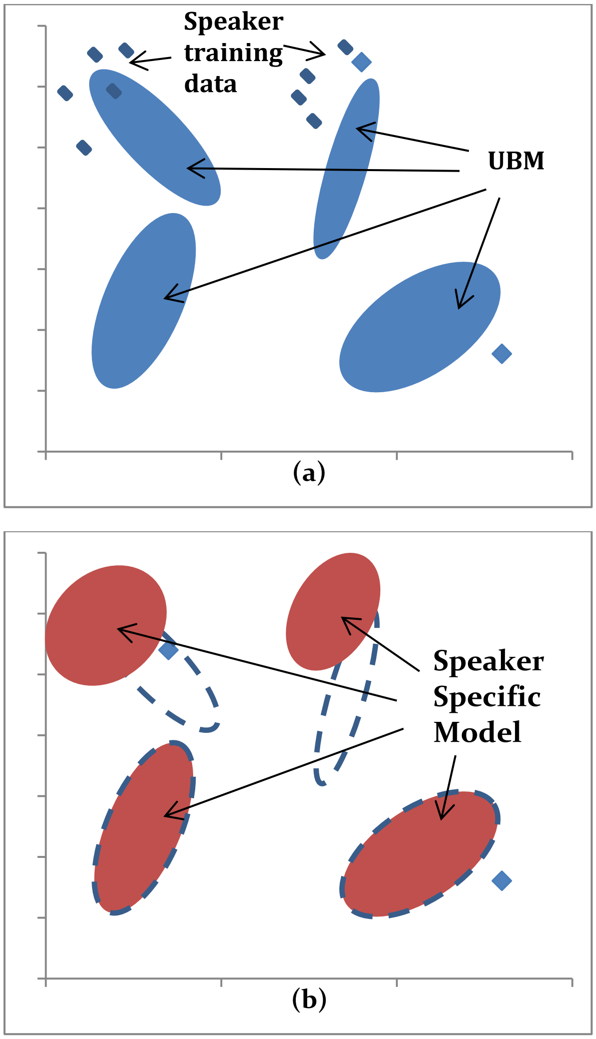


Figure 2-7 Graphical example of two of the early stages of the adaptation process. (a) The training vectors mapped to the UBM. (b) Derived adapted mixture parameters for the speaker-specific model.

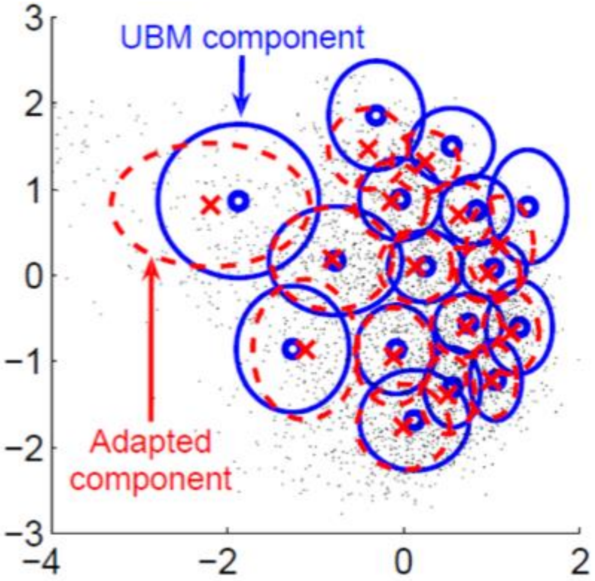


Figure 2-8 GMM Speaker model adaption using Maximum a posteriori (MAP) algorithm **[32]**

3 REPRESENTATION OF ACOUSTIC SIGNALS

3.1 Human Auditory Perception

The Human auditory system can perceive and distinguish audio signals based on three main areas, *pitch*, *loudness* and *timbre*. This auditory system goes through changes as we grow older [25]. Young people often have a better ability to hear high pitched sounds, but their hearing frequency range decreases about every 10 years. A pure tone can be described precisely by the frequency where it's generated and its intensity, but the pitch is not simple to define.

3.1.1 Pitch

Pitch can be defined as a perceived quantity related to frequency of vibration [25], because it is perceived, not all people can recognize a particular pitch value. The frequency at which the vocal cords vibrate is a function of their shape, tension and air flow over the vocal tract. Analyzing pitch is quite hard because it is very subjective across different people and it depends on their age group. Studies have shown that only 1 in 10,000 people can properly recognize absolute pitch values [33]. Typically we can agree that pure tones can be ordered in such a way, that relative to each other, one is 'higher' or 'lower' than the other [34].

Variations in pitch levels can represent different areas and can be subdivided. Typically when carrying a conversation, most of us don't use our whole pitch range, but understand that changing the impression of our voice can convey different messages such as emotion or subtle variation.

For example, we may recall an instance when an SMS text message we have sent may have conveyed a completely different message than intended, because it's hard to include paralinguistic messages in a SMS text message

[25]. Figures Figure 3-1 through Figure 3-4 show examples of these variations of the verb *try*.

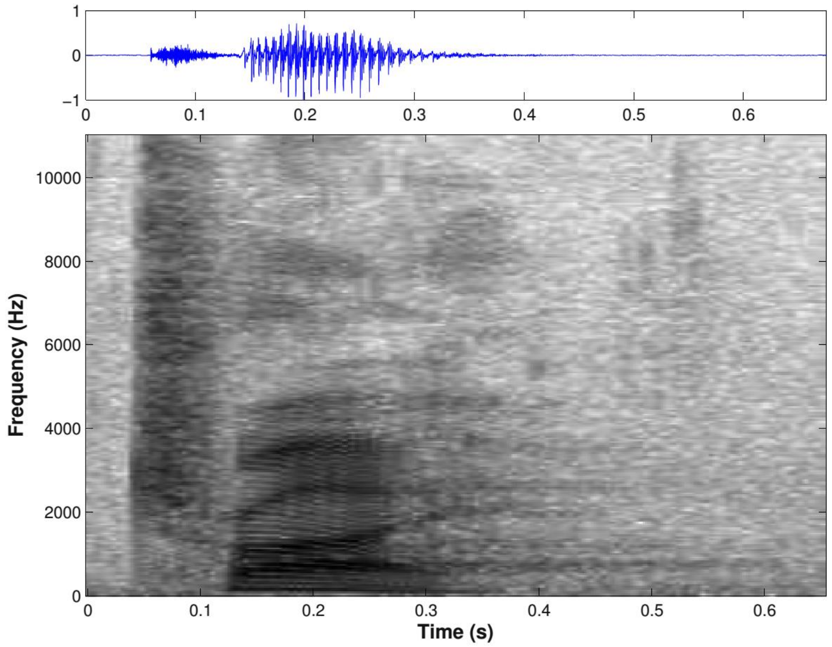


Figure 3-1– The verb [try] using a short, powerful expression (Decisive Imperative) **[25]**

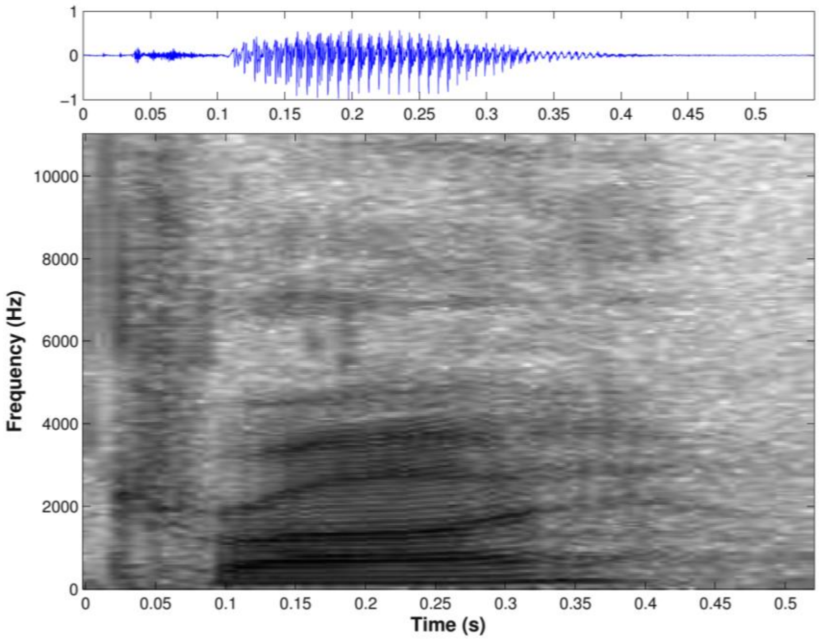


Figure 3-2– The verb [try] using strong and interrogative ending expression (Imperative & Interrogative) **[25]**

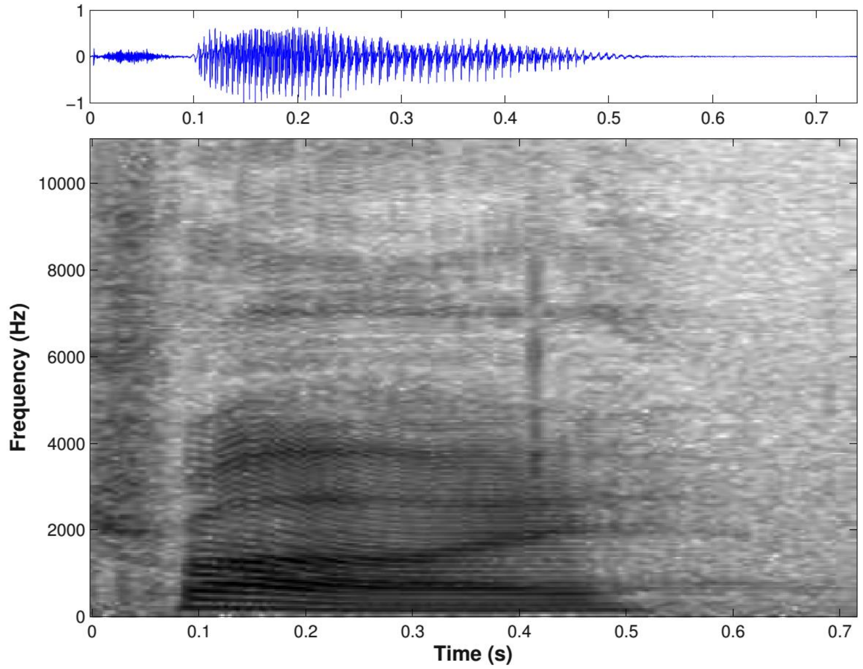


Figure 3-3– The verb [try] strong interrogative longer with higher pith level expression **[25]**

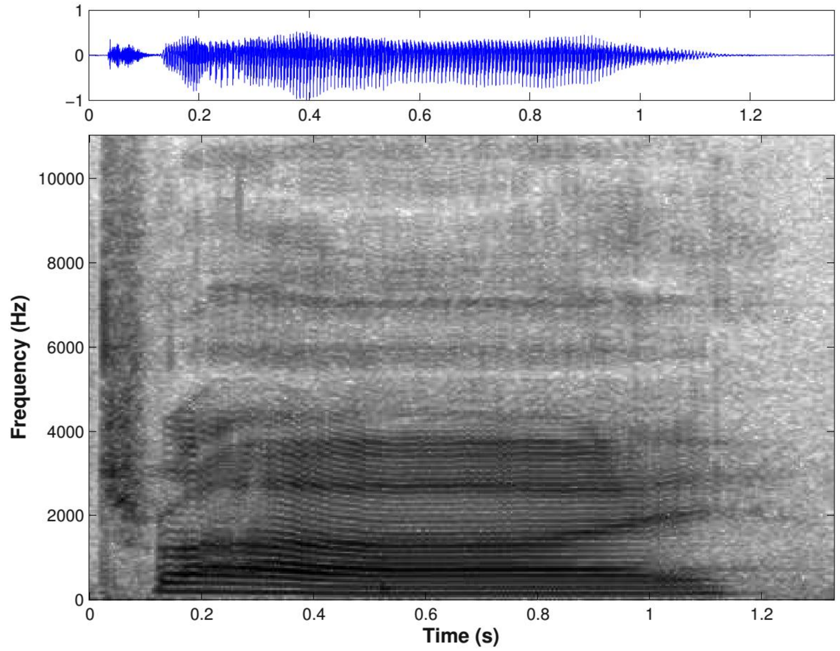


Figure 3-4– The verb [try] using a combination of pitch variations and a final drop **[25]**

3.1.1.1 Melody (Mel) Scale

In 1937 Stevens created a scale to represent the measurement unit of pitch: the *Mel Scale*. The Mel, short for Melody, scale is subjective as well, judged by different listeners to have equal distances given a reference tone [35]. The Melody (Mel) is a unit of pitch and is equal to one thousandth of the ($\wp$) of a pure tone, 40db above the listener threshold at a frequency of 1000 Hz [25]. This relation can be expressed better using Equation (3-1)

$$\wp = \frac{1000}{\log 2} \log\left(1 + \frac{f}{1000}\right) \qquad (3\text{-}1)$$

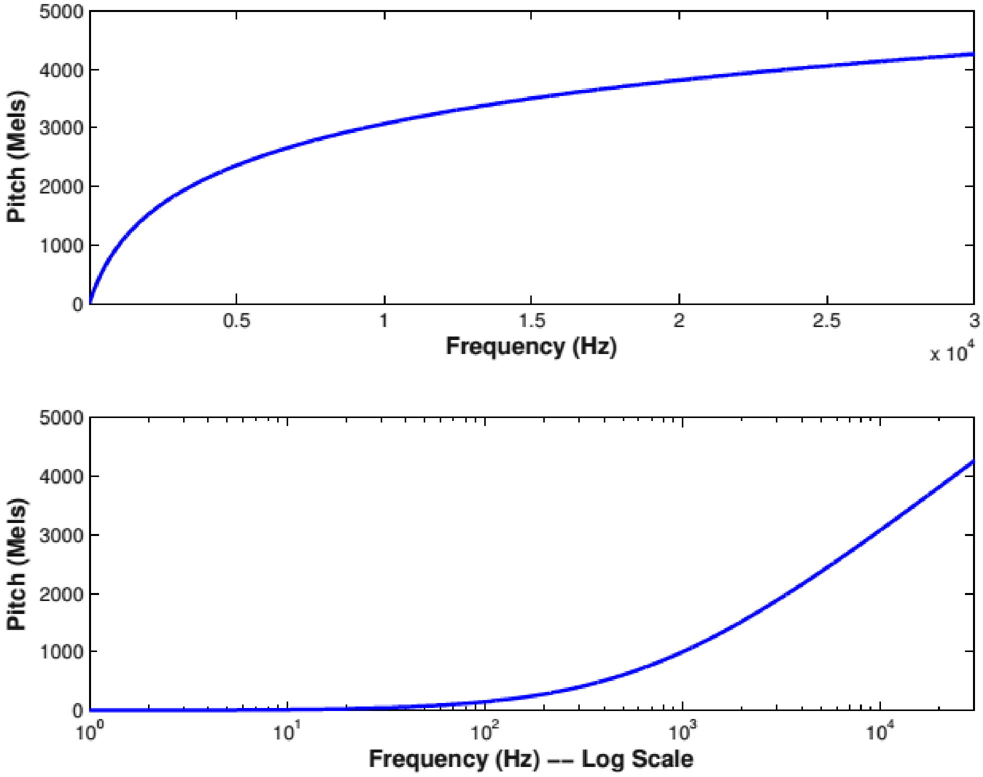


Figure 3-5 – Plot of Pitch vs Frequency (entire audible range) **[25]**

3.1.2 Loudness

Loudness is another perceived quantity that is a function of intensity and pitch. It can be defined as the intensity of vibration of the vocal cords over some duration of time at a particular pitch. One can think of this intensity as a measure of the power of the wave [25].

According to Beigi [25] the Intensity, $\boldsymbol{I}$, can be expressed as a relationship to the pressure differential, $\boldsymbol{P}$=$2 \times 10^{-5} \frac{N}{m^2}$ and the specific acoustic impedance of the sound medium, $\boldsymbol{\zeta}$=$413.21 \times 10^{-5} \frac{N_s}{m^3}$ according to the following equation,

$$\boldsymbol{I} = \frac{\boldsymbol{P}^2}{\boldsymbol{\zeta}} \qquad \textbf{(3-2)}$$

Therefore, (3-2) is the Intensity at 1000 Hz just enough for any person to hear the tone.

3.1.3 Timbre

Most accomplished musicians should be familiar with the concept of Timbre. The musical term *Timbre* is typically associated with the harmonic content and dynamic characteristics of the audio; in other words the frequency content of the source of the audio [36]. This kind of information reflects a specific characteristic for each speaker, which is essential in distinguishing them in speaker recognition.

4 MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

In order to create a robust speaker recognition system, you must have a mechanism to not only accurately represent the acoustic signals of a given speaker but it also has to be reliable. Fortunately a lot of research has been done on this area of signal acoustics. Research has led to a proven method to extract unique characteristics of speakers, the *Mel-Frequency Cepstral Coefficients* [37]. In this section we describe this feature extraction method widely used in Speaker Recognition as well.

The typical process for *feature extraction* can be seen on Figure 4-1 [38], with the assumption that it has been processed digitally and properly quantized.

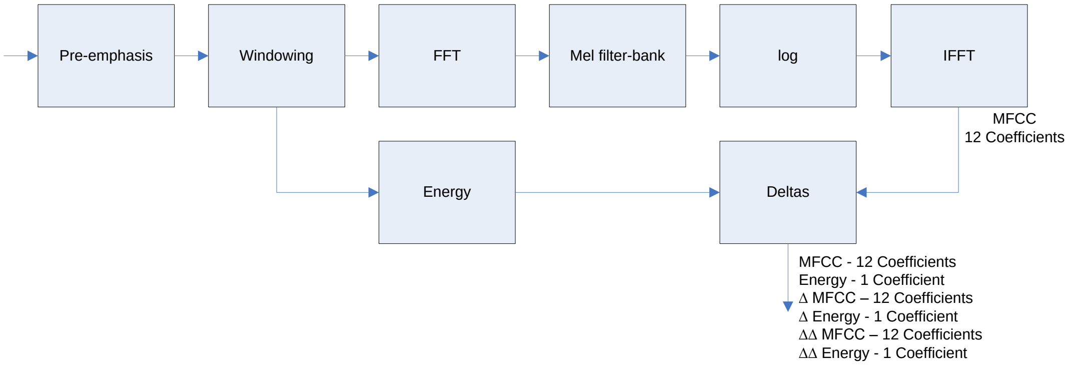


Figure 4-1 Feature extraction processing workflow [31]

4.1 Pre-emphasis

The speech signal is pre-emphasized to compensate for the spectral slope. The *spectral slope* is the tendency of natural audio signals to have less energy at high frequencies [39]. Pre-emphasis is actually a high pass filter operation to amplify the energy at high frequencies [34]. This actually reduces the difference in power components of the signal. This filter can be applied either in the frequency or time domain. In the time domain this filter can be defined using the following equation,

$$\boldsymbol{y_n = x_n - \propto x_{n-1}} \quad , 0.9 \le \propto \le 1.0 \qquad \textbf{(4-1)}$$

where the α is selected to be a value of *0.95* according to Combrinck and Botha [34]. The transfer function of this filter is given by,

$$\boldsymbol{H_p(z) = 1 - \propto z^{-1}} \qquad \textbf{(4-2)}$$

Figure 4-2 shows the power spectral density of a speech waveform before and after pre-emphasis. You can notice in the original signal the drop at higher frequencies compared to the pre-emphasized signal using α = 0.95, in which the power is better distributed across the relative frequencies. This comparison can also be seen using the spectrograms in Figure 4-3. Notice that the high frequencies are more prevalent in the pre-emphasized signal.

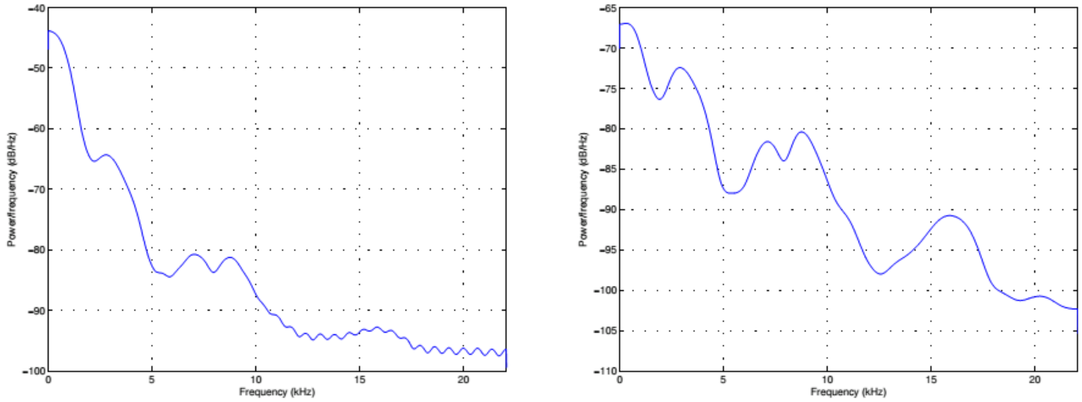


Figure 4-2 Power spectral density speech signal sampled at 44100 Hz before/after pre-emphasis [25]

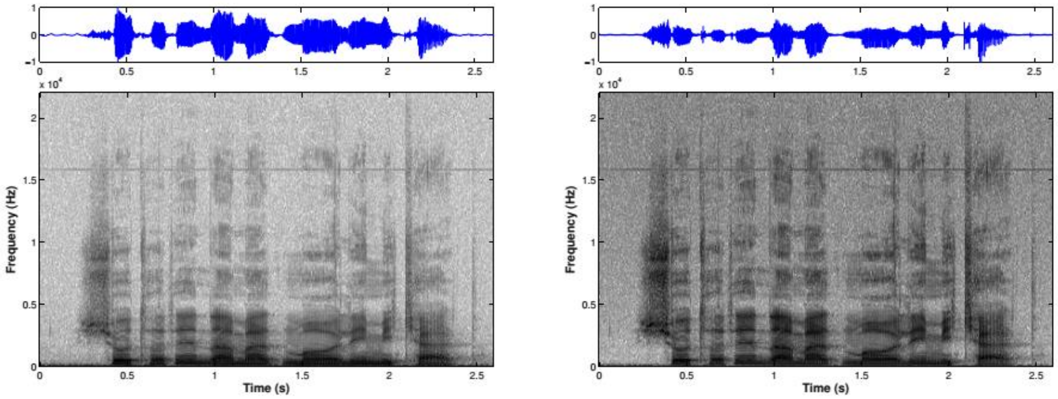


Figure 4-3 Spectrogram of a speech signal sampled at 44100 Hz before and after pre-emphasis [25]

4.2 Window

The windowing process is the one where we take the samples of the signal and multiply them by a window function. This is done to reduce any signal discontinuities [34], effectively slicing the signal into discrete segments [38]. A popular choice is the *Hamming window* because it prevents any sharp edges like rectangular windows. Figure 4-4 shows the window in the time and frequency domains. Equation (4-3) describes the Hamming window

$$\boldsymbol{w}[\boldsymbol{n}] = \begin{cases} 0.54 - 0.46\,\mathbf{cos}\,\dfrac{2\pi\boldsymbol{n}}{\boldsymbol{L}} & 0 \le \boldsymbol{n} \le \boldsymbol{L} - 1 \\ 0 & \end{cases} \tag{4-3}$$

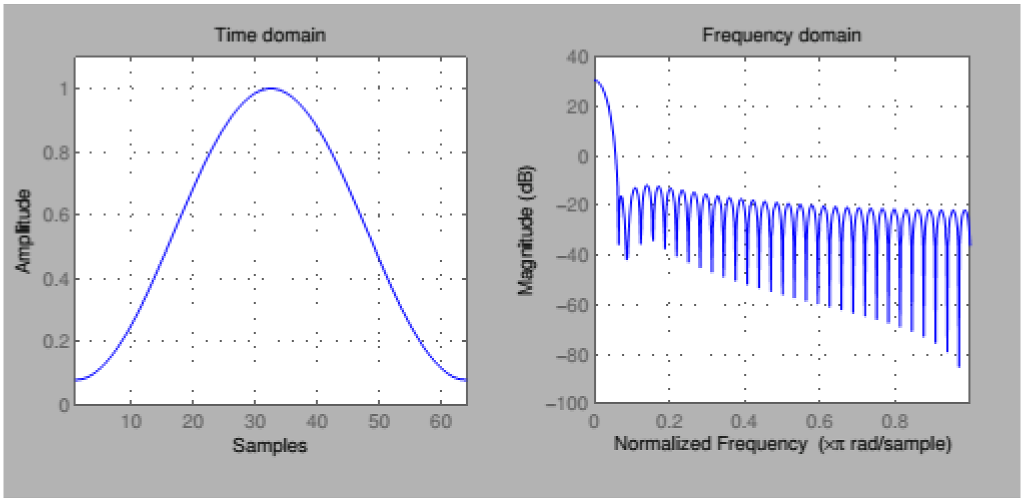


Figure 4-4 The Hamming window [25]

The effect of the window on a speech signal can be seen in Figure 4-5.

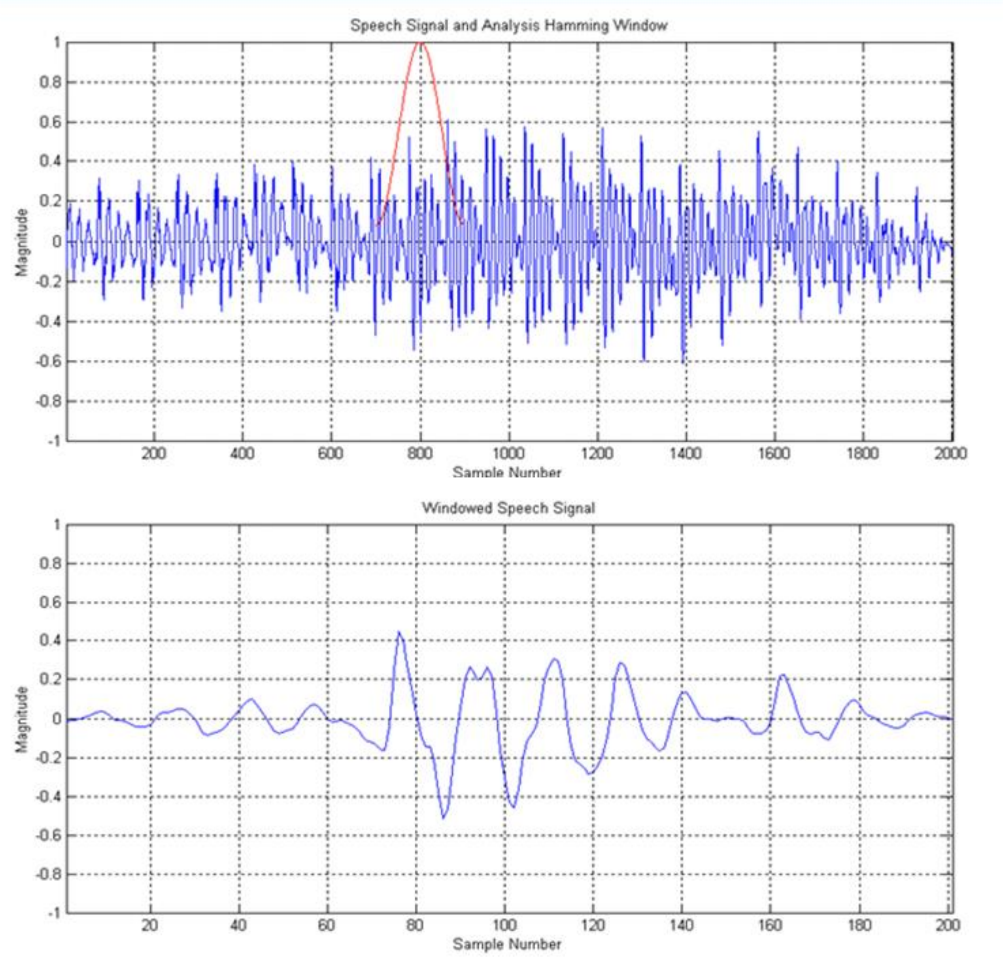


Figure 4-5 Original and Windowed Speech Signal [38]

4.3 DFT

The Discrete Fourier Transform (**DFT**) is one of the most widely used transforms especially in signal processing because it converts a sequence from the time domain to frequency and vice versa [40]. Therefore the Discrete Fourier Transform can be defined as,

$$H_k = \sum_{n=0}^{N-1} \boldsymbol{h}_n \boldsymbol{e}^{-i\frac{2\pi kn}{N}} \qquad \textbf{(4-4)}$$

A DFT can be seen in Figure 4-6 Sequence of N=10 samples DFT Figure 4-6, it shows a sequence of 10 samples. This DFT is then applied to the windowed signal. The result of this operation yields the magnitude and the phase representation of the signal, as can be seen on Figure 4-7.

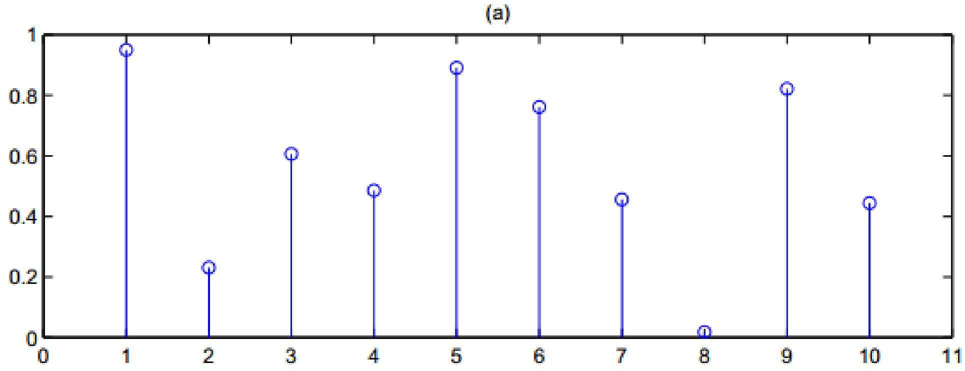


Figure 4-6 Sequence of N=10 samples DFT [41]

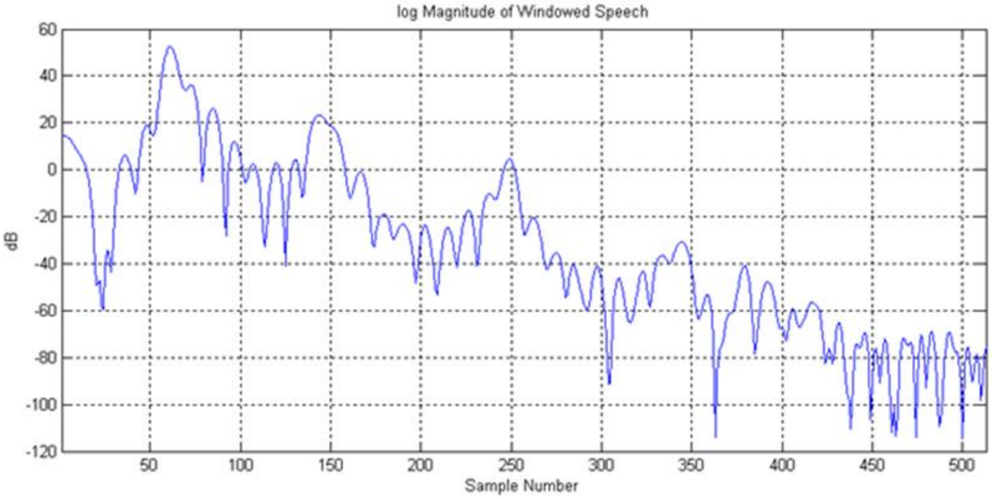


Figure 4-7 Magnitude of the windowed speech signal [38]

4.4 Mel Filter-Bank

As discussed earlier in section 3.1.1.1, the perception range of a human can be defined using the Mel Scale. The experiments of Zwicker [42] modeled the human auditory system using a 24-band filter-bank. The result of the previous stage, DFT spectrum, does not take into account that human hearing is less sensitive at frequencies higher than 1000 Hz. The DFT calculations only pertain to a linear frequency scale, therefore, we have to apply a process called *frequency warping*, in which the spectrum frequencies have to be converted to smaller numbers using the logarithmic Mel Scale.

In order to achieve this we can build a filter bank as presented in Figure 4-8 and effectively map the DFT frequency bin centers. This filtering is also known as the Mel-Spectrum [38] defined as follows,

$$\boldsymbol{mel}(\boldsymbol{f}) = \mathbf{1127}\,\mathbf{ln}(\mathbf{1}+\frac{\boldsymbol{f}}{\mathbf{700}}) \tag{4-5}$$

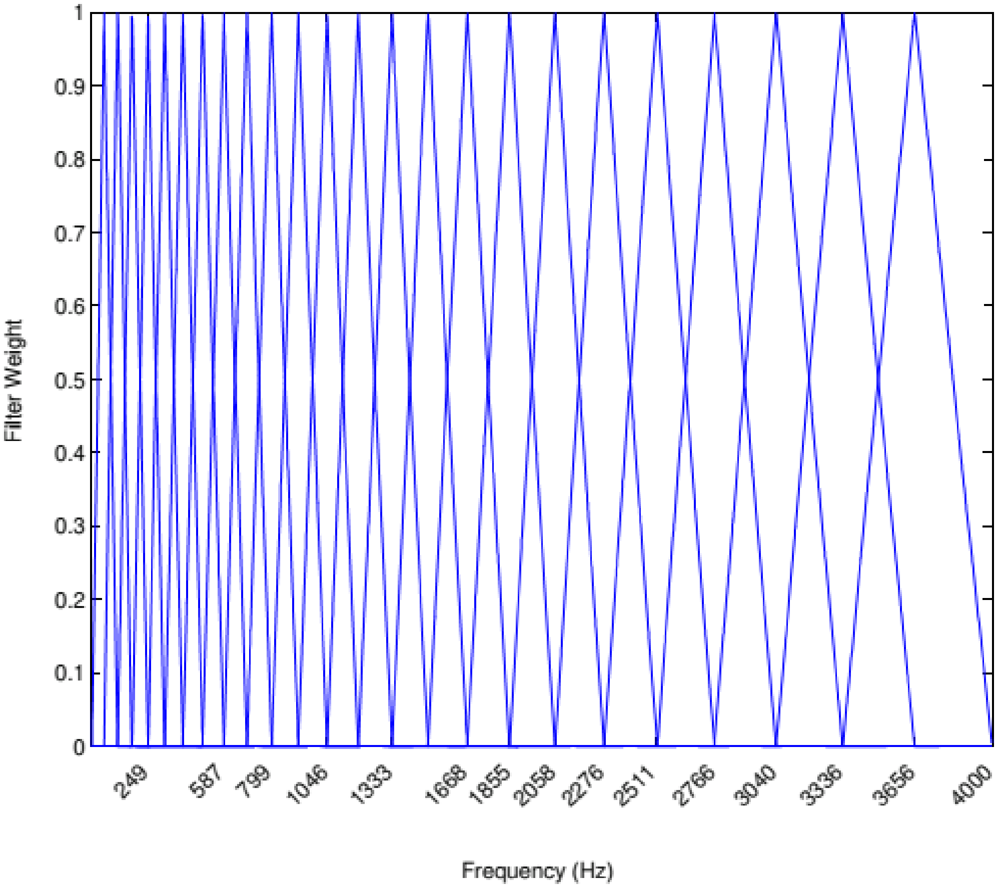


Figure 4-8 Shape of triangular Mel Filter banks for a 24-filter system sampled at 8000 Hz [25]

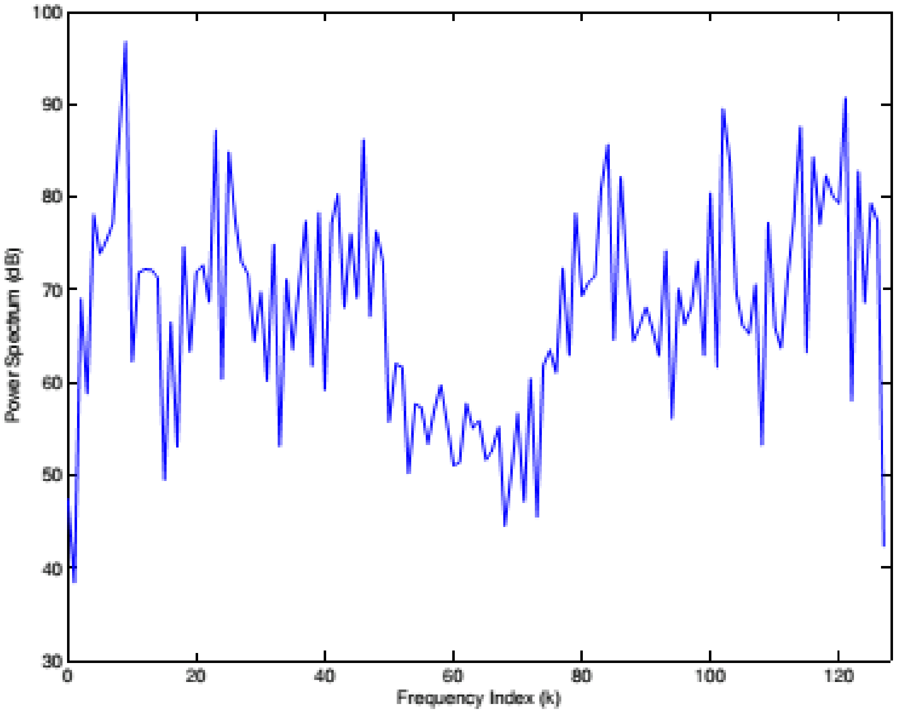


Figure 4-9 Power Spectrum of the frame to the left and to the right [25]

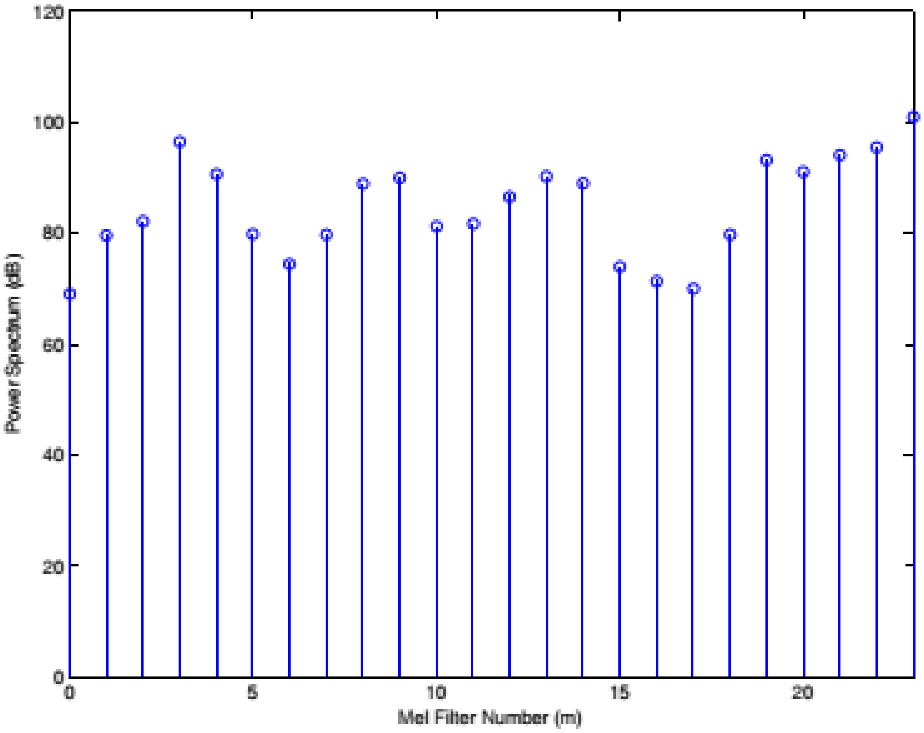
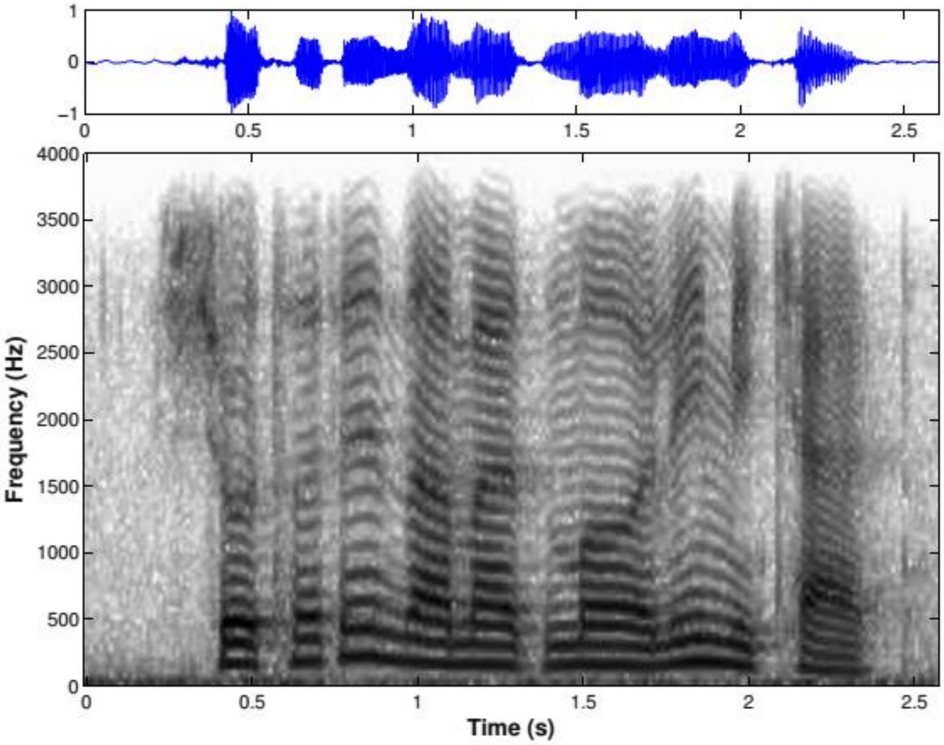
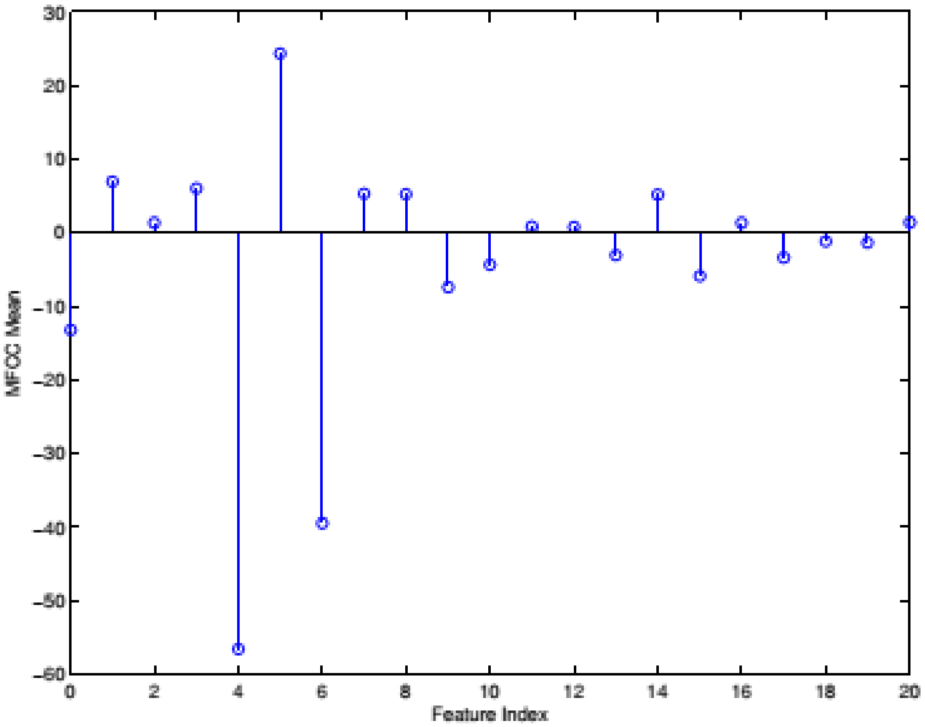


Figure 4-10 Mel-Spectrum, the Power Spectrum of the frame in the Frequency domain **[25]**

4.5 IDFT

The inverse Discrete Fourier Transform (IDFT) of the Mel-Spectrum is then computed, yielding the Mel-Frequency Cepstral Coefficients. The analysis of the signal in this new Cepstral domain proves to be beneficial given the fact of its inherent invariance toward linear spectral distortions. The first 12 values of the Cepstrum contain the meaningful information to provide unique characteristics of the waveform.


1
0
−1
0
0.5
1
1.5
2
2.5
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency (Hz)
Time (s)


30
20
10
0
-10
-20
-30
-40
-50
-60
MFCC Mean
0
2
4
6
8
10
12
14
16
18
20
Feature Index

4.6 Deltas

To capture frame to frame changes in the signal, the 1st and 2nd derivatives of the MFCC coefficients is calculated and also included.

Table 4-1 Total Number of Resulting Features in a Standard Speaker Vector

Feature Type	Count
Cepstral Coefficients	12
Delta Cepstral Coefficients	12
Double Delta Cepstral Coefficients	12
Energy Coefficient	1
Delta Energy Coefficient	1
Double Delta Energy Coefficient	1
Total	**39**

4.7 Energy

This stage happens simultaneously while the MFCC feature extraction is happening. The total energy of input frame is calculated.

5 GAMMATONE FREQUENCY CEPSTRAL COEFFICIENTS (GFCC)

One of the biggest problems in ASR is noise robustness. The sensitivity to additive noise is one of the major disadvantages of MFCC. The Gammatone Frequency Cepstral Coefficients (GFCC) are auditory feature based on a set of Gammatone Filter banks. A *Cochleagram* is a frequency-time representation of the signal and can be obtained from the output of the Gammatone filterbank. To compute the GFCC features a cochleagram is needed; the different stages of its computation have similarities with those of section 4, MFCC counterpart [43].

5.1 Gammatone Filter-Bank

The Gammatone filters are designed to simulate the process of the human auditory system. A Gammatone filter with a center frequency f_c can be defined as:

$$\boldsymbol{g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \varphi)} \qquad (5\text{-}1)$$

where $\boldsymbol{\varphi}$ is the phase but is usually set to zero, the constant a controls the gain and the order of the filter is defined by the value n which is typically set to a value less than 4 [43]. The factor b is defined as:

$$\boldsymbol{b = 25.17\left(\frac{4.37f_c}{1000} + 1\right)} \qquad (5\text{-}2)$$

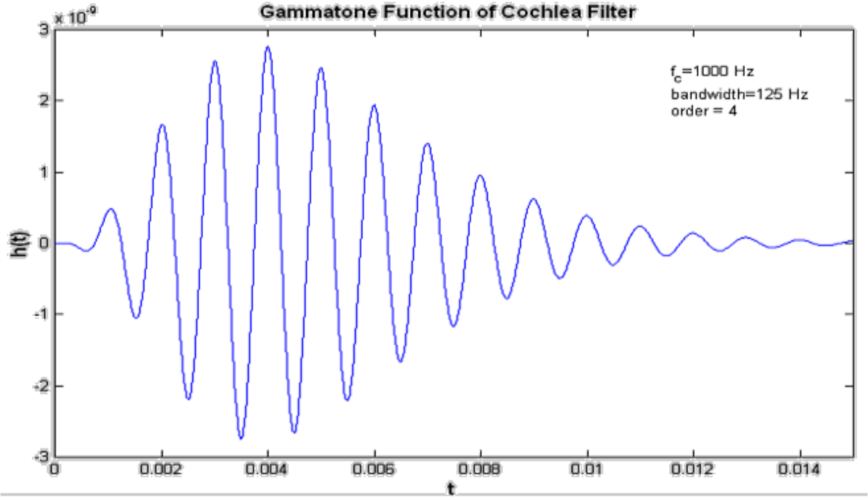


Figure 5-1 Impulse response of a Gammatone filter **[44]**

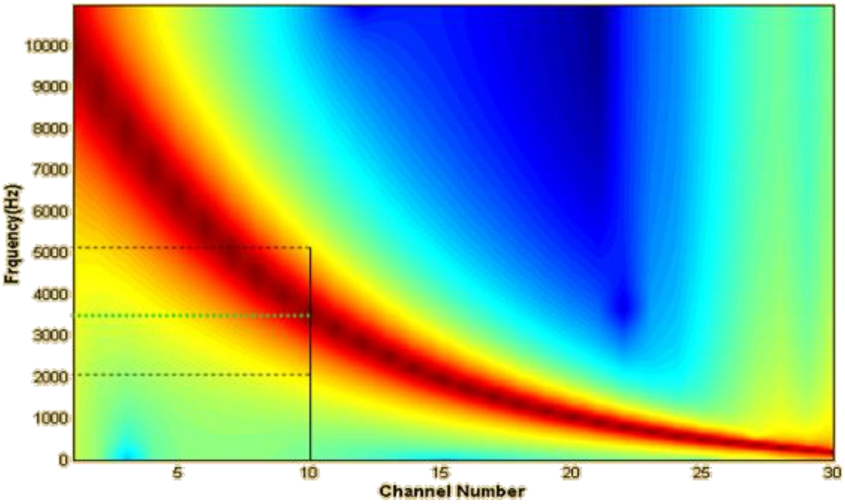


Figure 5-2 Gammatone filter characteristics **[44]**

To obtain a representation similar to an FFT based spectrogram a set of Gammatone filters, often referred as channels with different center frequencies, is used to create a Gammatone filter-bank. The frequency response of a 30 channel filterbank can be seen in Figure 5-3.

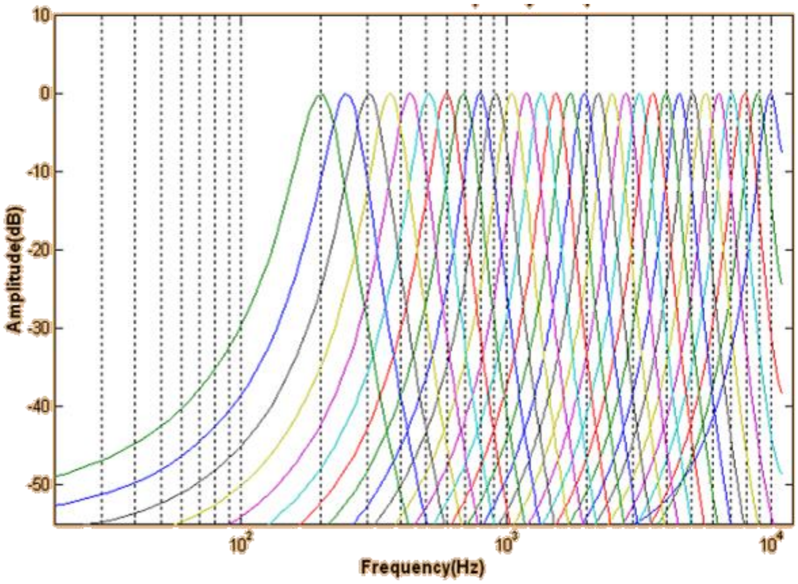


Figure 5-3 Frequency response of a 30-channel filterbank, 200-11025 Hz range **[44]**

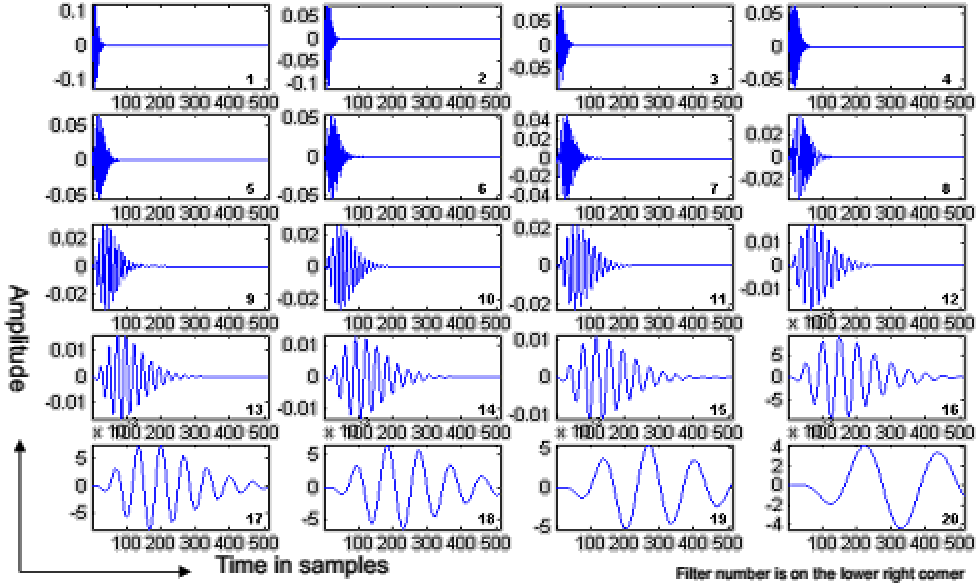


Figure 5-4 Impulse responses of individual filters of a 20 channel filter-bank **[44]**

5.2 Pre-emphasis

This step is similar to the pre-emphasis phase of the MFCC counterpart. It is employed to help reduce the dynamic range and to accentuate the frequency components that hold most of the key information needed from the speech signal.

Following the same idea from section 5.1 we define the pre-emphasis as a second order filter as follows,

$$\boldsymbol{H(z) = 1 + 4e^{-2\pi b/f_s}z^{-1} + e^{-2\pi b/f_s}z^{-2}} \tag{5-3}$$

Where f_s is the sampling frequency and b is the decay factor from (5-2)

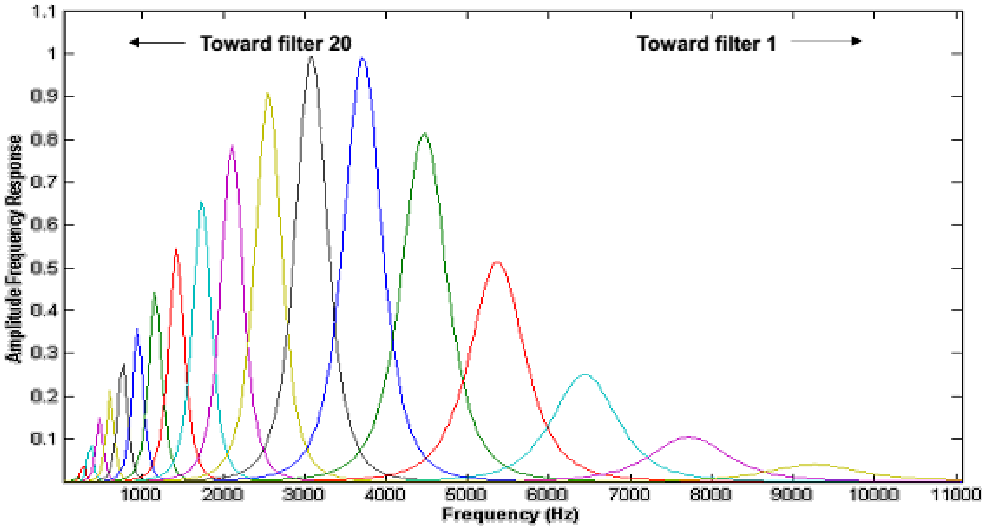


Figure 5-5 Gammatone filter output after applying pre-emphasis filter **[44]**

5.3 Window

Similar to the windowing process in section 4.2, GFCC needs a window applied to cover K points and shifts every L point for each frame. Each frame can be defined as $y(t;f_c(m))$, where f_c is the center frequency of the m-th filter. The resulting Cochleagram representation for each frame is computed averaging $y(t;f_c(m))$ across the window $t \epsilon (nL, nL + K)$ and is defined as follows [43] :

$$\bar{y}(n;m) = \frac{1}{K}\sum_{0}^{K-1} \gamma |y(nL + i; f_c(m))| \qquad \textbf{(5-4)}$$

where $\boldsymbol{\gamma}$ is a frequency dependent factor, and the other term represents the magnitude of a complex number. Aggregating $\bar{y}(n;m)$ across all of the channels yields

$$\bar{y}(n) = [\bar{y}(n;0), \ldots \bar{y}(n;M-1)]^T \qquad \textbf{(5-5)}$$

with M being the number of channels of the filter-bank. Typical values suggested by Wang & Xu [43] are K = 400, L=160 and M= 32 for 16 KH speech signal result in a 100 frames per second. The resulting matrix of this aggregation, $\bar{y}(n;m)$ is the Cochleagram.

5.4 DCT

The discrete cosine transform is then applied to obtain the uncorrelated cepstral coefficients. Similar to the MFCC operation, a log is also applied to the calculation.

The following equation describes this operation

$$g(n;u) = \left(\frac{2}{M}\right)^{0.5} \sum_{i=0}^{M-1} \left\{\frac{1}{3}\log\big(\bar{y}(n;i)\big)\cos\left[\frac{\pi u}{2M}(2i-1)\right]\right\} \qquad \textbf{(5-6)}$$

where typical u ranges are from 0 to 31. The first 12 components are then selected [43] resulting in a 12 dimensional GFCC feature:

$$g(n) = \ [g(n;0), \ldots g(n;11)]^T \qquad \textbf{(5-7)}$$

5.5 Deltas

To capture some of the temporal information, the GFCC is augmented with the 1st and 2nd order derivatives, bringing the total GFCC features to 36.

Table 5-1 Total Number of Resulting Features in a GFCC Vector

Feature Type	Count
GFCC Coefficients	12
Delta GFCC Coefficients	12
Double GFCC Cepstral Coefficients	12
Total	**36**

6 PRINCIPAL COMPONENT ANALYSIS

The basis of Principal Component Analysis (PCA), also known as *Karhunen-Loeve* [45], is a simple one: reducing the dimensionality of a data set. It transforms the spaces into a lower dimensional data space, a linear orthogonal transformation along the principal components of the space. It is very useful to analyze data and detect patterns more clearly. These *principal components* are a linear combination of the optimally-weighted observed variables. These optimum basis vectors are the eigenvectors of the covariance matrix of the distribution.

Sometimes data can be redundant, if we assume the set of features are somehow correlated, then we should be able to reduce the feature vectors without losing a lot of the information. PCA achieves this using *Eigensystem decomposition* generating an orthogonal transformation matrix that transforms the original feature vectors to a lower space [25]. This new space orders the principal components in terms of the variance of the corresponding dimension. The first principal component is the one that has the largest variance and the last few have the least variance, which in our case can be ignored.

The Figure 6-1 shows two principal component eigenvectors of some random data.

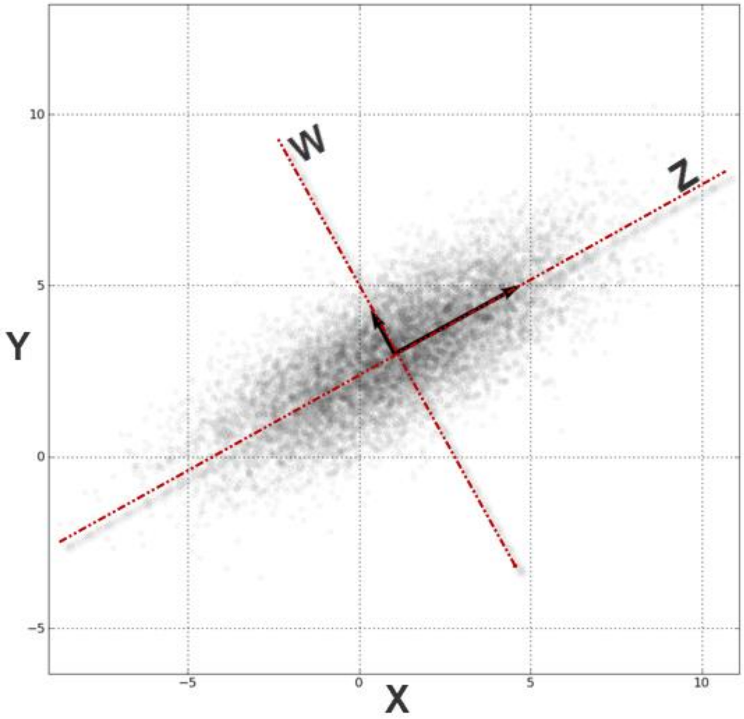


Figure 6-1 Two principal Components of some random data

7 SPEAKER RECOGNITION USING MFCC AND GFCC

Previous studies have shown the accuracy of MFCC [28] under low noise conditions and the robustness of GFCC [43] in noisy environments. It would be beneficial to incorporate the benefits of these two approaches, to reduce or eliminate their individual drawbacks.

7.1 Speaker Combined Feature Representation

In this section we will focus our attention on the problem of robustness and accuracy under noisy conditions. The strategy we are proposing allows us to combine the feature vector of MFCC and GFCC and use PCA to reduce the feature dimension and remove correlations. The front-end block diagram of the system is depicted on Figure 7-1. The system is subdivided into two different subsystems: MFCC and GFCC. Both systems will be running in parallel during the training and test phases. The output of these systems is aggregated and processed using statistical PCA.

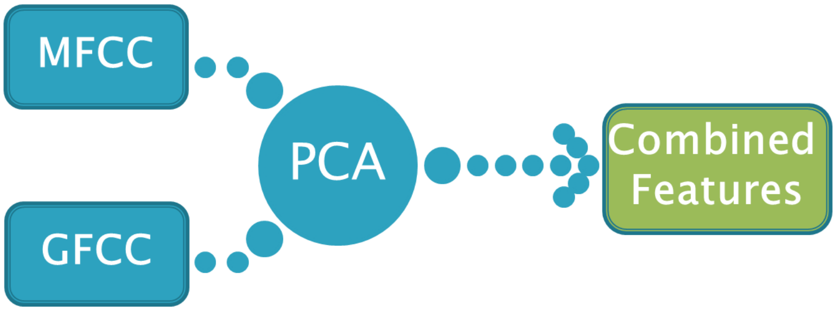


Figure 7-1 The combined feature representation front-end block diagram

The new feature set from the PCA after extraction is used to generate a characterized universal background model (UBM). The UBM is essentially a statistical background model from which we adapt the speaker model, following the conventional Gaussian mixture model (GMM) UBM framework [30]. The GMM is then trained using the background model from the complete set of speakers using the expectation-maximization (EM) algorithm. A speaker specific model is created and adapted from the UBM using maximum *a posteriori* (MAP) estimation. This signal flow can be seen on Figure 7-2,

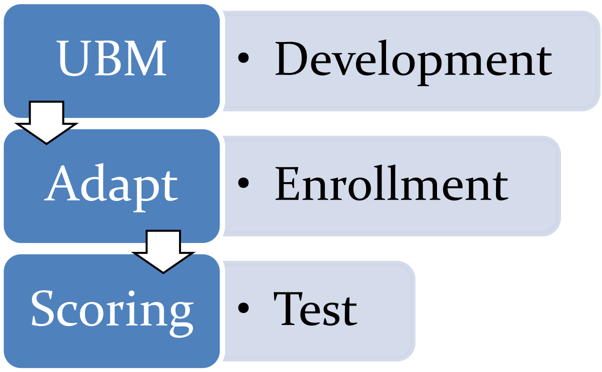


Figure 7-2 The GMM-UBM signal flow

8 TRAINING AND TEST DATA CORPUS

8.1 TIMIT

The **Texas Instruments and Massachusetts Institute of Technology (TIMIT)** corpus is designed to provide speech data for acoustic-phonetic studies and for speech recognition systems. TIMIT consists of data from 630 (192 female and 438 male) speakers in 8 dialects of U.S. English. Each speaker has 10 utterances, 2 *sa* sentences, 5 phonetically compact *sx* sentences and 3 phonetically *si* sentences. These sentences were carefully designed to contain a wide range of phonetic variability. Each utterance is recorded as a 16-bit waveform file sampled at 16 KHz.

8.2 ELSDSR

The **English Language Speech Database for Speaker Recognition** (**ELSDSR)** corpus is a dataset designed to provide speech data for the development and evaluation of ASR [46]. This corpus was developed by the faculty and graduate students from the Department of Informatics and Mathematical Modeling of the Technical University of Denmark. The ELSDR corpus consists of 22 speakers (10 female and 12 male) covering an age range from 24 to 63. Each speaker had to read an extensive and comprehensive message. Each utterance was recorded to a 16-bit PCM waveform with a sampling frequency of 16 KHz. The suggested training data for each speaker was created with seven paragraphs of text, which contained 11 sentences for a total of 154 utterances collected. The suggested test data was created with two sentences, 44 utterances. The

Table 8-1 shows the time duration for both training and test individually.

Table 8-1 ELSDR Duration of Reading Training and Test Material

No.	ID	Train (s)	Test (s)
Male			
1	MASM	81.2	20.9
2	MCBR	68.4	13.1
3	MFKC	91.6	15.8
4	MKBP	69.9	15.8
5	MLKH	76.8	14.7
6	MMLP	79.6	13.3
7	MMNA	73.1	10.9
8	MNHP	82.9	20.3
9	MOEW	88.0	23.4
10	MPRA	86.8	9.3
11	MREM	79.1	21.8
12	MTLS	66.2	14.05
Average		78.6	16.1
Female			
13	FAML	99.1	18.7
14	FDHH	77.3	12.7
15	FEAB	92.8	24.0
16	FHRO	86.6	21.2
17	FJAZ	79.2	18.0
18	FMEL	76.3	18.2
19	FMEV	99.1	24.1
20	FSLJ	80.2	18.4
21	FTEJ	102.9	15.8
22	FUAN	89.5	25.1
Average		88.3	19.6
Total		1826.6	389.55

9 CLASSIFICATION SCORING BASELINE

9.1 Experimental Setup

During the evaluation phase, each test segment is scored against the background model and a given speaker model to accept/reject the claim. The same set of tests is performed on both corpora.

The experiment extracts 39-dimensional MFCCs from a pre-emphasized speech signal, mean and variance normalization and writes them to disk in HTK format. The second stage extracts 36-dimensional GFCC's from the same speech signal and stores it to the disk in HTK format as well. The last stage uses the output of the MFCC and GFCC as the input to the PCA function. The first 30 principal components are used to reduce the dimensionality of the feature vectors; this new feature space is also saved in HTK format. To complete the experiment, the following steps are executed: UBM training, MAP adaptation, scoring of the verification trials, and computing the performance measures.

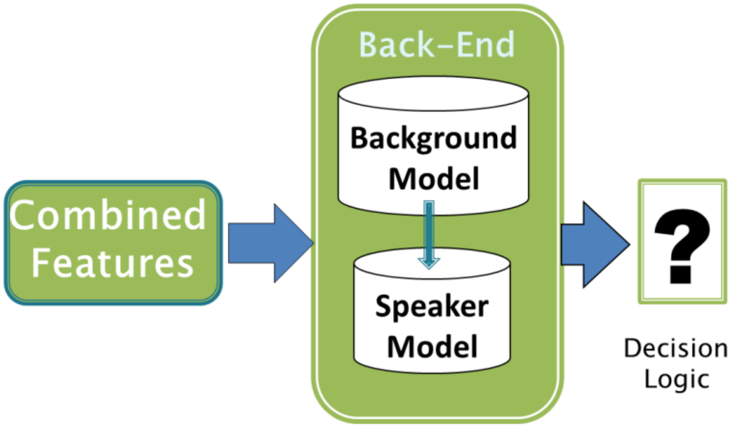


Figure 9-1 Block diagram of combined features experiment

For evaluating the performance of the new features in noise, the white Gaussian noise is added to the speech signal in different SNRs from -30dB to 0 dB, respectively. All model creation, training and test have been carried out using MSR Identity Toolbox v1.0 [47]**.** The TALSP [48] toolbox was used to generate the GFCC features and Voicebox [49] for MFCC's. The overall block diagram is shown in Figure 9-1.

9.1.1 MSR Identity Toolbox

The MSR Identity toolbox was developed by Microsoft Research as a MATLAB toolbox to help with speaker-recognition research [47]. It provides researchers with a collection of tools and a test bed to quickly build baseline systems for experiments. The toolbox provides paradigms for both GMM-UBM and i-vector. The toolbox provides some of the following capabilities: Feature normalization, GMM-UBM, i-vector-PLDA and EER & DET plot

9.1.2 UBM Training

From all of the TIMIT speakers 530 were selected for the background model training. The remaining 100 (30 female and 70 male) speakers are used for tests. The background model training phase uses all of the sentences from all 530 speakers. The speaker model training uses 9 out of 10 sentences per speaker, the last sentence is used for tests. Verification trials consist of all possible model-test combinations, making a total of 10,000 trials (100 target vs 9900 impostor trials).

For the ELSDSR background model training a similar approach was used, 18 (8 female and 10 male) speakers were selected. The remaining 4 speakers are used for the test trials. The verification trials consist of 16 trials (4 target vs 12 impostor trials).

The GMM was trained using 256 GMM components.

9.1.3 MAP Adaptation

This stage adapts the speaker specific GMM from the UBM using maximum *a posteriori* (MAP) estimation as discussed earlier in section 2.3.3. A MAP adaption relevance factor of 8.0 was used. The training data consisted of ten sentences per speaker for the TIMIT Corpus; only nine sentences were used for the speaker specific model. The ELSDSR Corpus consisted of nine sentences and only eight were used.

9.1.4 Scoring Verification Trials –

The verification scores for trials are computed as the log-likelihood ratio between the speaker models and the UBM given the test observations. The National Institute of Standards and Technology (NIST) has created a set of standard performance metrics to score ASR systems. The NIST Speaker Recognition Evaluation (SRE) has been performing a series of evaluations annually coordinated by the U.S. Department of Commerce since 1996.

9.1.4.1 Types of Errors

In statistical hypothesis testing there typically two types of errors, *false positives* and *false negatives*, often considered false alarms. A false positive is when a system incorrectly verifies an impostor as the target during the verification impostor trials. On the other hand a false negative is when the system determines the target as an impostor during the verification target trials. These types of errors are often referred to as false alarms and misses respectively.

The NIST evaluations have used a linear combination of the false alarm and miss error rates as its primary evaluation metric [50]. A decision cost function (DCF) is defined as

$$DCF = C_m \, x \, P_{m|t} \, x \, P_t \, x \, C_{fa} \, x \, P_{fa|i} \, x \, (1 - P_t) \qquad \textbf{(9-1)}$$

where C_m represent the cost of a miss, $P_{m|t}$ the prior probability of a miss given a target trial, P_t the prior probability of a target trial, C_{fa} the cost of a false alarm and $P_{fa|i}$ the prior probability of a false alarm given an impostor trial. Typical parameter values for the NIST evaluations are $C_m = 10$, $C_{fa} = 1$ and $P_t = 0.01$.

The NIST evaluations have also required the systems to produce a score along with the decision, where higher scores indicate greater likelihood that the correct decision is "true" [50]. A very informative way of presenting the system performance is a liner plot of both error rates on a normal scale, denoted by the NIST as the Detection Error Tradeoff (DET) curve [51].

The resulting curve is linear when the underlying error rates are normal. The Equal Error Rate (EER) is the critical operating area of the curve where the error rates (False Alarms and Misses) are equal.

9.2 Baseline Results

The performance baselines used for comparison are the individual features, against the combined feature set. The Table 9-1 and Figure 9-2 shows the summary of the EER achieved for each feature extraction technique. Figure 9-3 through Figure 9-7 shows the DET curves with all the features for the test trails at the appropriate SNR level.

Table 9-1 Summary of the Equal Error Rates (EER) and the Decision Cost Function (DCF) for the different SNR levels

SNR (dB)	EER % MFCC	EER % GFCC	EER % Combined	DCF
-30	49.929	25	25.303	10
-15	38.859	24	22.848	9.97
-10	27	22.879	17.121	9.49
-5	16	17.252	13.818	7.19
0	12	13.33	10.475	6.27

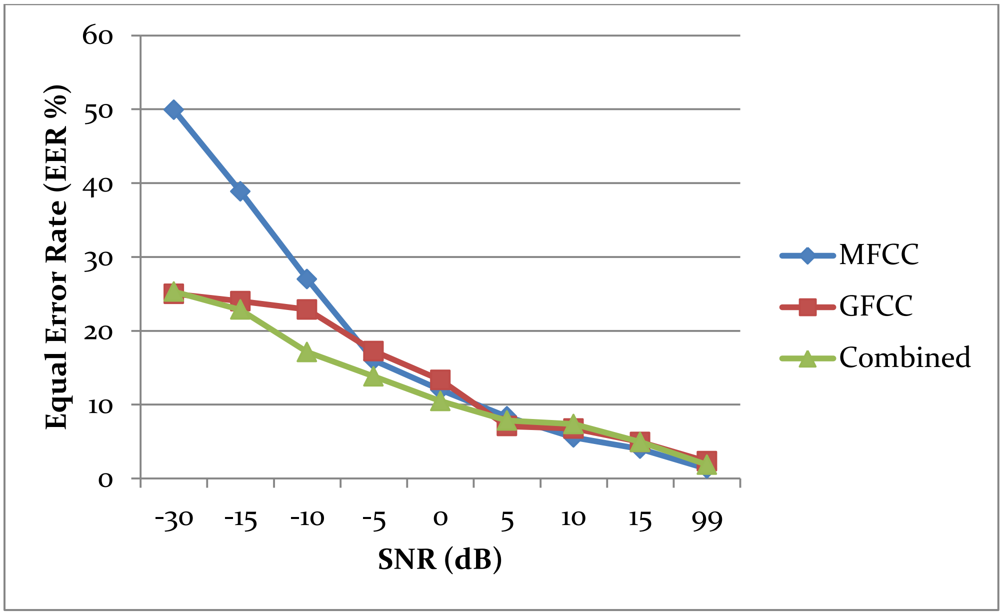


Figure 9-2 Final Total Equal Error Rates for the Test trials

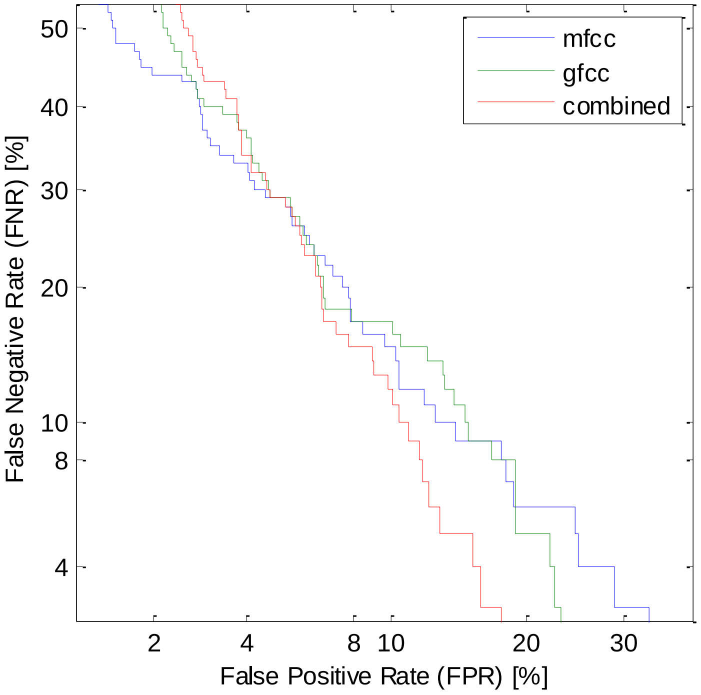


Figure 9-3 DET Curve using 0dB SNR level

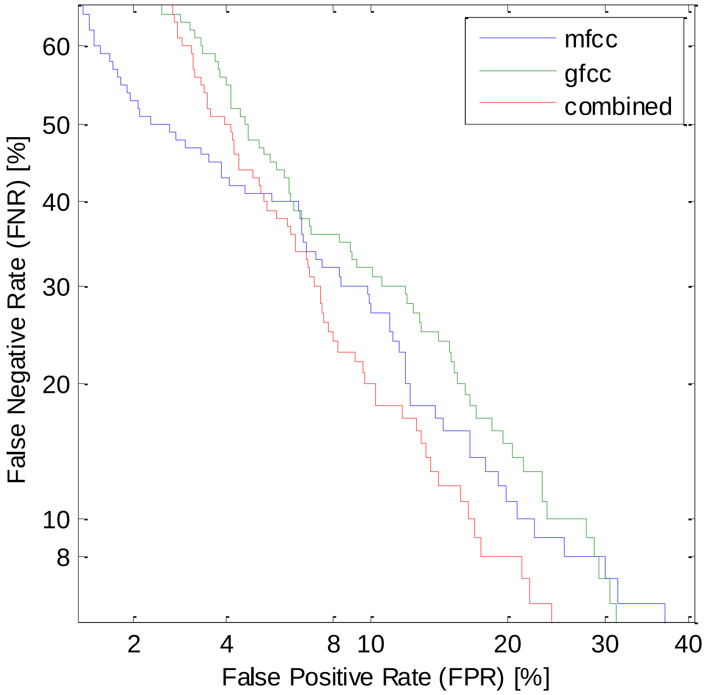


Figure 9-4 DET Curve using -5dB SNR level

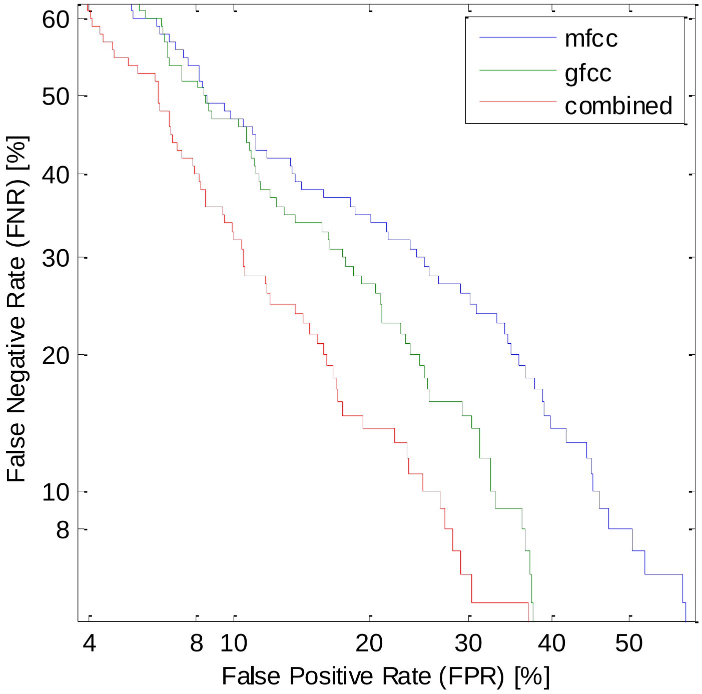


Figure 9-5 DET Curve using -10dB SNR level

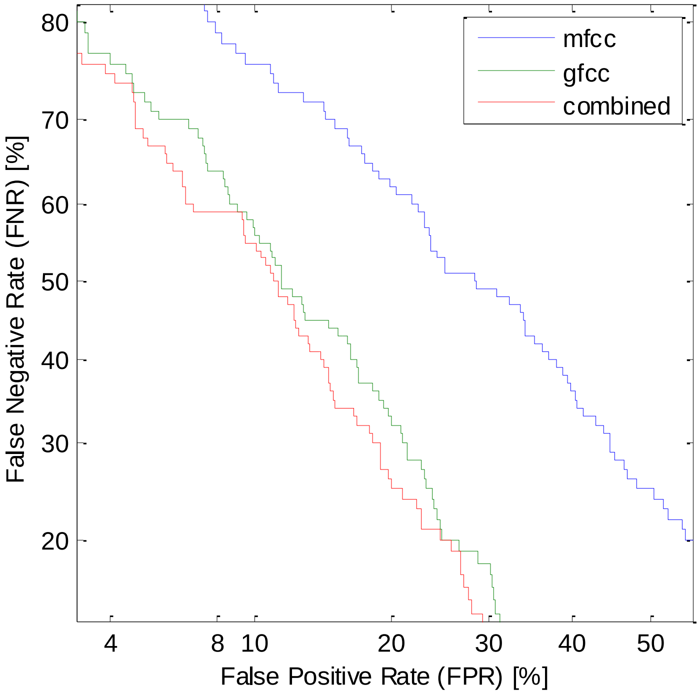


Figure 9-6 DET Curve using -15dB SNR level

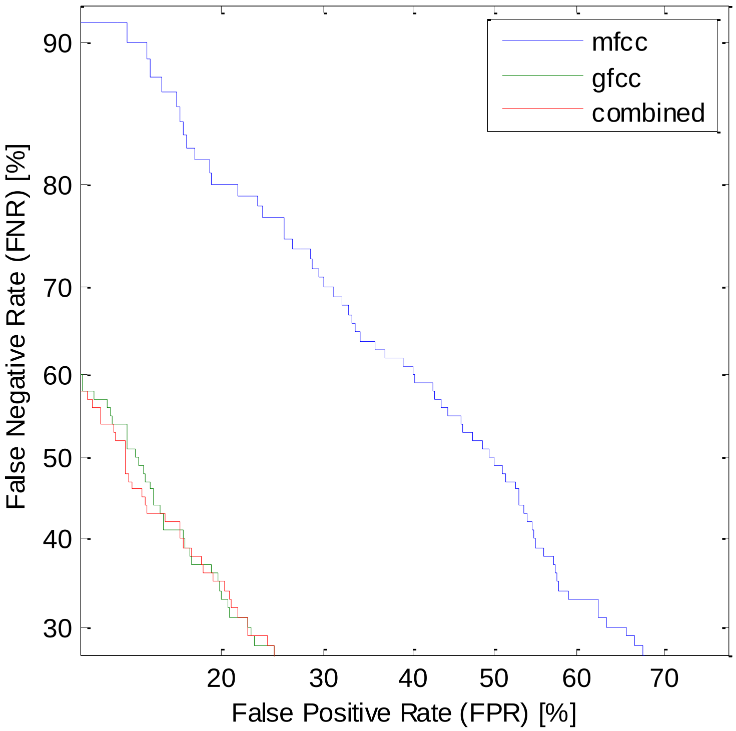


Figure 9-7 DET Curve using -30dB SNR level

10 CONCLUDING REMARKS

A combined approach for feature extraction has been presented and compared with MFCC and GFCC feature extractions algorithms.

The proposed combination feature methodology has shown satisfactory versatility and robustness under noisy conditions against the well-known TIMIT and the ELSDSR dataset. The final results in Table 10-1 shows that for the SNR levels tested overall there were significant improvement against the single feature counterparts. The highest improvement against MFCC was found at the -30dB range in which the EER improved 49%.

The results also show that the combined MFCC-GFCC is indeed a viable method to improve recognition rates at low SNR levels.

Table 10-1 Final Summary showing the EER improvement against the single features

SNR (dB)	% improvement against MFCC	% improvement against GFCC
-30	49.322	-1.21
-15	41.201	4.80
-10	36.59	25.166
-5	13.636	19.906
0	12.71	21.439

11 FUTURE WORK

In future work we will concentrate on improving efficiency by finding better methods of implicit/explicit feature combination. In addition we will investigate other types of features and incorporate new strategies to combine them. We can expand the studies even further by including speaker adaptive training and discriminative training. Finally, it should be noted that the equal error rates presented still are much better than the best combination results, i.e. the potential of system combination by far is not fully exploited, yet. Therefore, further research into improved model and/or system combination methods is due.

12 BIBLIOGRAPHY

[1] S. Graven and J. V. Browne, "Auditory Development in the Fetus and Infant," *Newborn & Infant Nursing Reviews,* pp. 187-193, December 2008.

[2] B. S. Kisilevsky, S. M.J., K. Lee, X. Xie, H. Huang, H. H. Ye, K. Zhang and Z. Wang, "Effects of Experience of Fetal Voice Recognition," *Psychological Science,* pp. 220-224, May 2003.

[3] B. H. Juang and L. R. Rabiner, "Automatic Speech Recognition - A Brief History of the Technology Development," Elsevier Encyclopedia of Language and Linguistics, 2005.

[4] J. Wayman, A. K. Jain, D. Maltoni and D. Maio, Biometric Systems: Technology, Design and Performance Evaluation, Springer, 2005.

[5] NSTC Biometrics, "Speaker Recognition," 7 August 2006. [Online]. Available: http://www.biometrics.gov/Documents/speakerrec.pdf. [Accessed March 2014].

[6] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *Journal of the Acoustical Society of America,* pp. 354-358, 1963.

[7] S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," *Acoustical Society of America,* pp. 2041-2047, 1964.

[8] K. P. Li, J. E. Dammann and W. D. Chapman, "Experimental Studies in Speaker Verification, Using an Adaptive System," *Acoustical Society of America,* p. 966, 1966.

[9] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *Acoustic Society of America,* pp. 637-655, 1971.

[10] G. R. Doddington, "A Method of Speaker Verification," *Acoustical Society of America,* pp. 49,139, 1971.

[11] W. Haberman and A. Fejfar, "Automatic ID of Personnel through Speaker and Signature Verification - System Description and Testing," in *1976 Carnahan Conference on Crime Countermeasur*, 1976.

[12] S. Furui, "40 Years of Progress in Automatic Speaker Recognition," in *Third International Conference on Biometrics*, Alghero, 2009.

[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, Signal Processing,* pp. 254-272, April 1981.

[14] J. M. Naik, L. Netsch and G. Doddington, "Speaker verification over long distance telephone lintes," in *International Conference on Acoustics, Speech and Signal Processing*, Glasgow, 1989.

[15] A. Rosenberg and F. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Compuer Speech and Language ,* pp. 1943-1957, 1987.

[16] S. Furui, "40 Years of Progress in Automatic Speaker Recognition," in *Advances in Biometrics*, Springer Berlin Heidelberg, 2009, pp. 1050-1059.

[17] T. Matsui and S. Furui, "Text-independent Speaker Recognition using Vocal Tract and Pitch Information," in *The First International Conference on Spoken Language Processing, ICSL*, Kobe, 1990.

[18] A. De la Torre, J. Segura, C. Benitez, J. Ramirez and A. J. Rubio, "Speech Recognition Under Noise Conditions: Compensation Methods," in *Robust Speech Recognition and Understanding*, I-Tech Education and Publishing, 2007, pp. 439-460.

[19] R. Rose, E. Hofstetter and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *Speech and Audio Processing, IEEE Transactions on (Volume:2 , Issue: 2),* pp. 245-257, 1994.

[20] "Speaker Recognition Evaluation," 5 March 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/sre.cfm.

[21] J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing,* pp. 430-451, 2004.

[22] M. Mclaren, R. Vogt, B. Baker and S. Sridharan, "A Comparison of Session Variability Compensation Techniques for SVM-Based Speaker Recognition," in *8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, 2007.

[23] A. Petri, J.-F. Bonastre, D. Matrouf, F. Capman and B. Ravera, "A Comparison of Session Variability Compensation Techniques for SVM-Based Speaker Recognition," in *8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, 2007.

[24] M.-C. Cheung, M.-W. Mak and S.-Y. Kung, "A two-level fusion approach to multimodal biometric verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[25] H. Biegi, Fundamentals of Speaker Recognition, Yorktown Heights: Springer, 2011.

[26] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[27] P. Saikia, D. Bora, A. F. Syiemlieh, C. S. K., S. Majumder and P. K. Dutta, "Real Time Speaker Recognition System using PCA and ICA," in *Michael Faraday IET India Summit*, Kolkata, India, 2012.

[28] M. Zamalloa, L. Rodríguez, M. Peñagarikano, G. Bordel and J. Uribe, "Improving robustness in open set speaker identification," in *Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008.

[29] M. Liu, T. Huang and Z. Z., "Robust Local Scoring Function for Text-Independent Speaker Verification," in *Proc. International Conference on Pattern Recognition (ICPR)*, Hong Kong, 2006.

[30] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian," in *Digital Signal Processing 10*, Lexington, Massachusetts, 2000.

[31] V. Kepuska, "Automatic Speech Recognition Class Slides," 18 November 2009. [Online]. Available: http://my.fit.edu/~vkepuska/ece5527/Ch5-Automatic%20Speech%20Recognition.pptx. [Accessed 12 February 2010].

[32] B. V. Srinivasan, Scalable Learning for Geostatistics and Speaker Recognition, College Park,Maryland: University of Maryland, 2011.

[33] T. Rossing, The Science of Sound, 3rd edn, Addison Wesley, 2001.

[34] H. Combrinck and E. C. Botha, "On The Mel-scaled Cepstrum," Dept of Electrical and Electronic Engineering, University of Pretoria, Pretoria, 1996.

[35] S. S. a. V. J. a. N. E. B. Stevens, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America,* vol. 8, pp. 185-190, 1937.

[36] A. J. M. Houtsma, "Pitch and timbre: Definition, meaning and use," *Journal of New Music Research,* vol. 26, no. 2, pp. 104-115, 1997.

[37] X. Huang, A. Acero and H.-W. Hon, Spoken Language Processing, A guide to theory, algorithm and system development, Prentice Hall PTR, 2001.

[38] A. G. Kunkle, Sequence Scoring Experiments Using the TIMIT Corpus and the HTK Recognition Framework, Melbourne: FIT, 2010.

[39] D. B. Fry, The Physics of Speech, Cambridge: Cambridge University Press, 1996.

[40] K. R. Rao, D. N. Kim and J. J. Hwang, Fast Fourier Transform: Algorithms and Applications, London, New York: Springer, 2010.

[41] S. Roberts, "Signal Processing and Filter Design Lecture Notes," 2003.

[42] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenz-gruppen)," *Journal of the Acoustical Society of America,* pp. 248-249, 1961.

[43] J. Qi, D. Wang, J. Xu and J. Tejedor, "Bottleneck Features based on Gammatone Frequency Cepstral Coefficients," in *Proc. of Interspeech 2013*, Lyon, France, 2013.

[44] W. H. Abdulla, "Auditory Based Feature Vectors for Speech Recognition Systems," The University of Auckland, New Zealand, 2012.

[45] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed, Orlando, FL: Academic Press, 1990.

[46] L. Feng and L. K. Hansen, *A NEW DATABASE FOR SPEAKER RECOGNITION,* Kongens Lyngby, Denmark: Technical University of Denmark, 2005.

[47] S. O. Sadjadi, M. Slaney and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," *IEEE SLTC Newsletter,* November 2013.

[48] X. Zhao, Y. Shao and D. Wang, "CASA-based Robust Speaker Identification," *IEEE Trans. on Audio, Speech and Language Processing,* vol. 20, no. 5, p. 1608, 2012.

[49] M. Brookes, "Voicebox: Speech processing toolbox for matlab," Imperial College, 2006. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[50] A. F. Martin, Speaker Databases and Evaluation, Gaithersburg, Maryland: National Institute of Standards and Technology, 2007.

[51] A. F. Martin, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997.

[52] Y.-W. Chen and C.-J. Lin, "Combining SVM's with Various Feature Selection Strategies," Taipei.