8-23-2016

Robust speaker recognition in the presence of speech coding distortion

Robert Walter Mudrosky
Rowan University

ROBUST SPEAKER RECOGNITION IN THE PRESENCE OF SPEECH CODING DISTORTION

by

Robert W. Mudrowsky

A Thesis

Submitted to the
Department of Electrical and Computer Engineering
College of Engineering
In partial fulfillment of the requirement
For the degree of
Master of Science in Electrical and Computer Engineering
at
Rowan University
August 10, 2016

Thesis Chair: Ravi P. Ramachandran, Ph.D.

Acknowledgments

I would like to thank Dr. Ravi Ramachandran for his confidence in me and affording me the opportunity to conduct this research and continue my education. I would also like to thank Dr. Umashanger Thayasivam, Dr. Linda Head, and Dr. John Schmalzel for their assistance and guidance in completing this research endeavor.

This work was supported by the National Science Foundation under grant DUE-1122296.

Abstract

Robert Mudrowsky
ROBUST SPEAKER RECOGNITION IN THE PRESENCE OF SPEECH CODING DISTORTION
2015-2016
Ravi P. Ramachandran, Ph.D.
Master of Science in Electrical and Computer Engineering

For wireless remote access security, forensics, border control and surveillance applications, there is an emerging need for biometric speaker recognition systems to be robust to speech coding distortion. This thesis examines the robustness issue for three coders, namely, the ITU-T 6.3 kilobits per second (kbps) G.723.1, the ITU-T 8 kbps G.729 and the 12.2 kbps 3GPP GSM-AMR coder. Both speaker identification (SI) and speaker verification (SV) systems are considered and use a Gaussian mixture model (GMM) classifier. The systems are trained on clean speech and tested on the decoded speech. To mitigate the performance loss due to mismatched training and testing conditions, four robust features, two enhancement approaches and feature (SI) and score (SV) based fusion strategies are implemented.

The first proposed novel enhancement method is feature compensation based on the affine transform and is used to map the features from the test scenario to the train scenario. The second is the McCree signal enhancement approach based on the spectral envelope information. A detailed two-way analysis of variance (ANOVA) supplemented with a multiple comparison test is performed in order to show statistical significance in application of these enhancement methods.

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

Table of Contents (Continued)

List of Figures

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

1.1 Statement of the Problem

The main objective in the design of any speaker recognition system is to maximize performance in regards to correctly identifying or verifying a given speaker for any test condition. The quality of speech passed through a speaker recognition system will have an effect on overall system performance. The degradation of this speech quality is apparent in many forms of additive noise which include echo, latency, packet loss, packet delay variation, and distortion originating from the speech coder [1][2]. Distortion introduced by the speech coder will degrade the speech quality which will reduce system performance. The examination of distortion originating from the speech coder will be the main focus of this study. A GMM-UBM (Gaussian Mixture Model-Universal Background Model) speaker recognition system is implemented for both speaker identification (SI) and speaker verification (SV) to investigate the problem of speech coder distortion. In this thesis, the term speaker recognition is generic and refers to speaker identification and/or speaker verification. Training of the SI and SV systems is done on clean speech. The testing phase is done on the decoded speech which is the clean speech passed through the speech coder and then, decoded.

1.2 Motivation

This study will examine three contemporary speech coders of various bitrates. The speech coders used are G729 and G723.1 from the ITU standards (International Telecommunications Union) as well as GSM AMR (Groupe Spécial Mobile Adaptive

Multi-rate codec) from the 3GPP (3rd Generation Partnership Project). The G.729 coder which is used primarily in VoIP (Voice over Internet Protocol) applications and uses a bit rate of 8 kbit/s [3][6]. The G723.1 coder is used in VoIP multimedia applications and uses a bit rate of 6.3 kbit/s [4][5]. The GSM AMR coder is a variable bitrate coder in which the bit rate of 12.2 kbits/s will be exclusively used in this study. GSM AMR is used primarily in mobile communication technologies [3][7]. These selections allow for a varied sampling of speech coders in current use. Each coder uses a different bit rate. The effect of the bit rate with regards to speech coding distortion will be investigated. Speaker recognition performance as a function of bit rate is investigated by simulating these three coders.

1.3 Objective of Thesis

The objectives of this thesis are:

1. To improve the performance of a speaker recognition system by reducing the effect of speech coder distortion.
2. To implement a GMM-UBM based system.
3. To implement feature enhancement by applying the Affine transform
4. To implement signal enhancement by applying the McCree method.
5. To combine feature and signal enhancement.
6. To implement post-processing fusion techniques to further augment performance.
7. To determine the optimal set of system parameters for the implementation of a speaker recognition system. These parameters include the number of Gaussian

mixtures, the speech features used, the type of enhancement method and the fusion strategy.

8. To apply statistical techniques to compare the different approaches to determine statistical significance.

1.4 Thesis Focus and Organization

The focus of this thesis is the implementation and analysis of a GMM-UBM based speaker recognition system designed to mitigate the effects of speech coding distortion and to improve overall system performance using feature and signal enhancement.

The first chapter is an introduction to the problem of speech coding distortion as well as a description of the purpose of this thesis.

The second chapter provides a background of the speech coding standards used, the training and testing parameters, a description of the features, a complete description of GMM-UBM system parameters, enhancement methods and fusion strategies.

The third chapter explains the design approach of the GMM-UBM speaker recognition systems and a detailed explanation of the experimental procedure for both SI and SV systems.

The fourth chapter contains the results and findings related to the GMM-UBM speaker recognition systems. The effectiveness of fusion strategies as well as analyses to determine statistical significance will be discussed.

The fifth chapter summarizes and lists the conclusions and successes of the thesis. Recommendations for potential future work and considerations are discussed as well.

Chapter 2

Background

This chapter contains a complete review of all the aspects related to the design of the speaker recognition systems for this thesis. The parameters of the narrow-band speech coders used in the experimentation are discussed. A comprehensive description of the feature extraction methods and related features are also discussed.

A discussion of the characteristics of the Gaussian Mixture Model (GMM) using a universal background model (UBM) speaker recognition system is provided. An explanation of maximum a-posteriori estimation (MAP) as well as the use of expectation maximization (EM) as it relates to the UBM is presented.

Two types of speaker recognition systems will be examined. An explanation of a speaker identification (SI) system and a speaker verification (SV) system as well as their respective performance metrics will be discussed.

The usage of enhancement methods and their variations, which are the primary contribution of this thesis, will be discussed. An explanation of the McCree method of signal enhancement and the affine transform which allows for feature enhancement will be examined. Various fusion methods to further augment speaker recognition system performance will also be discussed. A statistical analysis will also be performed in order to prove statistical significance. This includes a two-way analysis of variance (ANOVA) and a t-test.

2.1 Narrow-Band Speech Coding

The speech coders covered in this study operate using narrow-band audio channels which range from 300-3.4 kHz using a sampling frequency of 8 kHz [1]. This convention does not cover the entire human vocal range but it still allows for adequate intelligibility of speech. Preserving the intelligibility of speech is one of the primary goals of any speech coding algorithm. The three speech coders that used in this thesis adhere to these basic principles.

The coders under investigation provide a current sampling of contemporary speech compression methods. The relationship between system performance and the various bit rates of the coders will be examined.

2.1.1 G723.1. The G.723.1 speech coder is also an ITU standard used primarily for low bandwidth VoIP applications. There are two bit rates utilized by this speech coder. This thesis makes use of the 6.3 kbit/s bit rate option which employs a fixed frame size of 24 bytes per 30 ms frame. The G.723.1 speech coder uses multi-pulse linear predictive coding with maximum likelihood quantization (MPC-MLQ) algorithm [1][4][5].

2.1.2 G.729. The G.729 speech coder is an ITU standard used in wireless communication as well as VoIP applications where the conservation of bandwidth is a principal requirement. It operates at a fixed bit rate of 8 kbits/s and fixed frame size of 10 bytes per 10 ms frame. The G.729 speech coder uses a code-excited linear prediction algorithm (CELP) [1][6].

2.1.3 GSM-AMR. The GSM-AMR speech coder is a multi-rate speech coder which is a standard governed by the 3GPP (3rd Generation Partnership Project) primarily used in mobile phone applications. There are eight bit rates to choose from for this coder. This thesis will examine the 12.2 kbits/s bit rate selection that uses a fixed frame size of 244 bits per 20 ms frame. The GSM-AMR speech coder uses a CELP algorithm [3][7].

2.2 Features

Four feature sets are used in this thesis. The features are as follows: linear predictive cepstrum (CEP), adaptive component weighting weighted cepstrum (ACW), postfilter cepstrum (PST), and mel-frequency cepstral coefficients (MFCC). Linear predictive (LP) analysis is used for the CEP, ACW, and PST features [9][10]. The feature extraction process for MFCC is based on the filter bank processing of the Fourier transform of the speech followed by cepstral analysis using the discrete cosine transform (DCT) [2][19]. Energy thresholding is implemented in order to ensure that only frames that contain sufficient speech information are used when calculating the feature vectors.

2.2.1 Linear prediction. As stated above, the feature extraction process for CEP, ACW, and PST is accomplished by use of linear predictive (LP) analysis. Linear predictive analysis is based on the idea that a speech sample is a weighted linear combination of p previous samples which results in a set of weights labeled a_k [8].

The equation is given as:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n)$$

(2.1)

where *s(n)* is the speech signal and *e(n)* is the error or LP residual. The weights correspond to the coefficients of a non-recursive filter given as:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = \prod_{k=1}^{p} (1 - f_k z^{-1})$$

(2.2)

where f_k for $1 \leq k \leq p$ represents the zeros of *A(z)*. The calculation of the LP coefficients a_k is based on the minimizing the weighted mean squared-error E_{mse} on a segment of speech comprising of *N* samples. The weighting is accomplished by applying a Hamming window to the segment of speech. Finding a_k by minimization of the E_{mse} is accomplished by an autocorrelation analysis and solving a system of linear equations using the Levinson-Durbin algorithm. Using this algorithm assures minimum phase of *A(z)* [9].

The all-pole LP transfer function is given as:

$$H(z) = \frac{1}{A(z)} = \prod_{k=1}^{p} \frac{1}{1 - f_k z^{-1}} = \sum_{k=1}^{p} \frac{r_k}{1 - f_k z^{-1}}$$

(2.3)

where r_k represents the residues and f_k represents the poles of *H(z)*. The poles being represented as:

$$f_k = \sigma_k e^{j\omega k}, \quad k = 1,2, \dots, p$$

(2.4)

where ω_k is the k^{th} center frequency and σ_k is the magnitude of the poles that fall in the range of (0,1).

The causal impulse response is given as:

$$h(n) = \sum_{k=1}^{p} r_k f_k^n = \sum_{k=1}^{p} r_k \sigma_k^n \, e^{j\omega_k n}$$

(2.5)

Since *A(z)* is guaranteed to be minimum phase the CEP, ACW, and PST features are causal (exist only for quefrencies n ≥ 0) [9].

2.2.2 Linear predictive cepstrum feature (CEP). For a system function *P(z)*, the cepstrum is generally defined as the inverse z-transform of log[*P(z)*] [9] given as:

$$C(z) = \log \mathrm{P(z)} = \sum_{n} c_p(n) z^{-n}$$

(2.6)

A pole zero transfer function *P(z)* is given as:

$$P(z) = \frac{U(z)}{V(z)} = \frac{\prod_{k=1}^{u}(1 - u_k z^{-1})}{\prod_{k=1}^{u}(1 - u_k z^{-1})}$$

(2.7)

If *P(z)* is minimum phase, the cepstrum can be calculated by a recursion based on the polynomial coefficients or by taking into consideration the polynomial roots v_k and u_k given as:

$$cp(n) = \frac{1}{n}\sum_{k=1}^{v} v_k^n - \frac{1}{n}\sum_{k=1}^{u} u_k^n$$

(2.8)

where $n > 0$.

In the case of the linear prediction filter *A(z)*, the cepstrum corresponding to 1/*A(z)* or equivalently the inverse z-transform of log[1/*A(z)*] is referred to as the LP cepstrum and is denoted by $c_{lp}(n)$. The CEP feature is $c_{lp}(n)$ and can be efficiently and recursively calculated (without root finding) from the predictor coefficients a_n [9]as:

$$c_{LP}(n) = a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) c_{LP}(i) a_{n-i}$$

(2.9)

2.2.3 Adaptive component weighting (ACW). The ACW cepstrum is obtained by first performing a partial fraction expansion of the LP function *1/A(z)* which is shown as:

$$\frac{1}{A(z)} = \sum_{k=1}^{p} \frac{\lim_{z \to f_k} \left[\frac{1 - f_k z^{-1}}{A(z)}\right]}{1 - f_k z^{-1}} = \sum_{k=1}^{p} \frac{r_k}{1 - f_k z^{-1}}$$

(2.10)

where f_k are the poles of *A(z)* and r_k are the corresponding residues. The variations of r_k are removed by setting $r_k = 1$ for every *k*. Therefore, the corresponding transfer function is a pole-zero type of the following form:

$$\frac{N(z)}{A(z)} = \sum_{k=1}^{p} \frac{1}{1 - f_k z^{-1}}$$

$$\frac{N(z)}{A(z)} = \frac{1}{A(z)} \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - f_k z^{-1})$$

$$\frac{N(z)}{A(z)} = p\left[\frac{1-\sum_{k=1}^{p-1} b_k z^{-k}}{1-\sum_{k=1}^{p} a_k z^{-k}}\right]$$

(2.11)

It has been shown in [10] that *N(z)* is minimum phase by recognizing that a circle that encloses all of the zeros of a polynomial also encloses all of the zeros of its derivative. Standard polynomial root finding does not need to be applied and *N(z)* can be easily calculated from A(z) as shown in [10]. The ACW feature is determined by computing the cepstrum of *N(z)/A(z)* by a recursion based on the polynomial coefficients of *N(z)* and *A(z)* [9].

2.2.4 Postfilter cepstrum (PST). The postfilter is obtained from *A(z)* and its transfer function is given as:

$$H_{pst}(z) = \frac{A\left(\frac{z}{\beta}\right)}{A\left(\frac{z}{\alpha}\right)}$$

(2.12)

where $0 < \beta < \alpha \leq 1$. The cepstrum $H_{pst}(z)$ is the postfilter cepstrum (PST/PFL) which is equivalent to weighting the LP cepstrum [9] shown as:

$$cpst(n) = clp(n)[\alpha^n - \beta^n]$$

(2.13)

where $\alpha = 1.0$ and $\beta = 0.9$

2.2.5 Mel-frequency cepstral coefficients (MFCC). Unlike the other features used in this thesis, the mel-frequency cepstrum coefficients (MFCC) feature extraction method is not based on LP analysis. Instead, it is computed by the filter bank processing of the Discrete Fourier transform (DFT) of the speech followed by a cepstral analysis of the discrete cosine transform (DCT). The magnitude of the DFT is logarithmically smoothed using a mel spaced filter bank. The DCT of the filter bank outputs yield the MFCC which is a basically a compact representation of the spectrum of the speech [2][19].

2.2.6 Delta feature. In order to better capture transitional information between frames, a 12-dimensional delta feature is computed for the four features for each frame. A delta feature uses a frame span of five (current frame plus look ahead and behind two frames) in order to derive first derivative information [11]. A delta feature can be computed using the following equation:

$$\Delta f_k = \frac{\sum_{n=-m}^{m} n f_{k+i}}{\sum_{n=-m}^{m} n^2}$$

(2.14)

where f_k is a feature vector at frame k and m = 2 corresponds to a frame span of 5. To obtain second derivative information the delta feature at frame k (Δf_k) is used as an input to once again calculate the above equation. Concatenation of the first and second derivative of the feature vector results in a 36 dimensional vector [11].

2.3 Speaker Recognition Systems

A speaker identification system (SI) and speaker verification system (SV) are considered in this thesis. A SI system determines the closest identity of a test utterance based on all available speaker models which is a 1:N problem. A SV system determines if the test speaker's claimed identity matches only the target speaker model which is a 1:1 problem.

Two different performance metrics are used. The SI system performance is measured by the identification success rate (ISR) in which the total number of correct identifications is divided by the total number of test trials. The SV system performance is measured using the equal error rate (EER) which is the operating point on the receiver operating characteristic (ROC) where the false accept rate (FAR) equals the false reject rate (FRR). A false acceptation is when the test speaker in question is accepted by the SV system when it actually should be rejected. The number of false acceptations divided by the total number of acceptances equals the FAR [3]. A false rejection is when the test speaker in question is rejected by the SV system when it actually should be accepted. The number of false rejections divided by the total number of rejections equals the FRR [3]. A ROC curve is a plot that depicts the FAR against the FRR. Both speaker recognition systems make use of a GMM-UBM classifier which is described in the following sections.

2.3.1 Gaussian mixture model (GMM). A Gaussian Mixture Model classifier is used as the basis of both speaker recognition systems. A GMM speaker model is described as a conditional probability density expressed as a linear combination of Gaussian densities [11] shown as:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i \, p_i(x)$$

(2.15)

where **x** is a *D*-dimensional feature vector, and w_i are the mixture weights which satisfies $\sum w_i = 1$ for $i = 1$ to **M** where **M** is the number of Gaussian Mixtures. The density $p_i(x)$ is given as:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T {\Sigma_i}^{-1}(x-\mu_i)\right\}$$

(2.16)

where μ_i is a D x 1 mean vector and Σ_i is a D x D covariance matrix. The parameters are denoted as $\lambda = \{w_i,\ \mu_i,\ \Sigma_i\}$ [11] [12].

2.3.2 Expectation maximization (EM). Expectation maximization (EM) is an iterative technique for maximum likelihood estimation (MLE). The maximum likelihood estimates of λ are obtained using EM [17][18]. There are two steps involved in each iteration of the EM algorithm. The first step is to compute the posterior probability given the current model and the second step is to update the model using the equations for the weights, means, and covariances. These two steps are iterated until the desired

convergence criteria have been satisfied. This refines the GMM parameters which increases the likelihood that the estimated model is closer to the observed feature vectors [1][3][12][17][18].

2.3.3 Universal background model (UBM). A Universal Background Model (UBM) is an alternative speaker model which consists of speakers pooled together that represent the expected speech characteristics of the speakers that will be enrolled in the SI and SV systems. It can be thought as one very large GMM that represents the impostor space [12]. The selected speech from speakers for the UBM is from a different partition of the TIMIT database then that of the speech from speakers that are enrolled in the SI and SV systems. For every mixture, the weights, means, and variances are computed using the EM algorithm from $i = 1$ to M where M is the number of mixtures [20]. This is repeated for all of the utterances used (10) for all of the speakers (168) to create the UBM

Once the UBM is created it is then adapted to develop the individual speaker models. The UBM serves as the initial condition in the training phase for the MAP adaptation of the GMM models for all speakers. There are two ways in which to perform the MAP adaptation of the GMM models. The first way is to use all of the statistics which include the weights, means, and variances and the second way is to use the means only. It has been shown in [12] that use of only the means is not sufficiently different when compared to using all three of the statistics. The GMM models are also computed for the number of mixtures for every training utterance (8) for each speaker (90 total). Ideally this computation for each mixture will gradually make the speaker model more robust.

Once training is complete the UBM is no longer used in regards to the SI system. When testing the SI system a test utterance is input and the feature vectors are created. A log likelihood based score for every speaker GMM model is then calculated. The identity of the speaker is specified as the largest score out of all of the compared GMM models.

The UBM has an essential role in regards to the testing of the SV system. A test utterance is input and feature vectors are created as in the SI system. However there are two sets of scores for the SV system. The true score is computed as the difference between the single target speaker model score and the score for the UBM. The true score is required to calculate the FRR [12]. The target speaker is in reality the claimed speaker and is compared to their actual GMM speaker model as shown in the following figure.

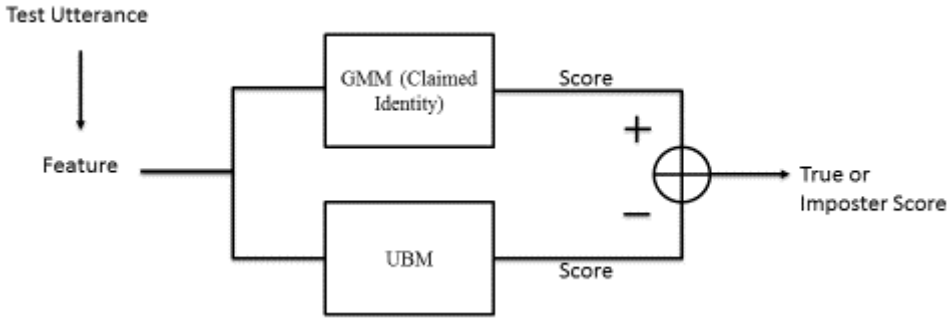

Figure 2.1. True/imposter score calculation

The imposter score is computed in the same way as the true score except that the target speaker is not actually the claimed speaker so it is not compared to their correct GMM speaker model. The imposter score is required to calculate the FAR. Once both scores are calculated then the FAR and FRR can be calculated which then allows for the EER to be calculated which is the performance metric for the SV system [3][12][13][14].

2.4 Enhancement Techniques

There are two pre-processing enhancement techniques utilized in this thesis. The principal contribution of this thesis is the application of the affine transform as a form of feature enhancement. The other technique is a form of signal enhancement. There are also unique fusion strategies implemented for both the SI and SV systems.

2.4.1 Affine transform. The affine transform enables feature enhancement by mapping a feature vector derived from the test speech to another feature vector in the region of the D-dimensional space occupied by the clean speech training vectors. This allows for a more consistent match between training and testing conditions which enhances the feature in question by compensating for this distortion [11]. The affine transform is given as:

$$y = Ax + b$$

(2.17)

where **A** is a **p** by **p** matrix and **y**, ***x*** and **b** are column vectors of dimension *p*. Expansion of equation 2.*17* results in:

$$\begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ \cdot \\ \cdot \\ \cdot \\ y(p) \end{bmatrix} = \begin{bmatrix} a_1^T \\ a_2^T \\ a_1^T \\ \cdot \\ \cdot \\ \cdot \\ a_p^T \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \cdot \\ \cdot \\ \cdot \\ x(p) \end{bmatrix} + \begin{bmatrix} b(1) \\ b(2) \\ b(3) \\ \cdot \\ \cdot \\ \cdot \\ b(p) \end{bmatrix}$$

(2.18)

Where $\mathbf{a_m^T}$ is the row vector corresponding to the *m*th row of *A*. Parameters **A** and **b** are determined using only the training data. The feature vector for the *i*th frame of the training speech is labeled as $\mathbf{y^{(i)}}$**.** The feature vector for the *i*th frame of the training speech with coder distortion is labeled as $\mathbf{x^{(i)}}$**.** A total of *N* sets of vectors are collected from $\mathbf{y^{(i)}}$ and $\mathbf{x^{(i)}}$ and a squared error function [11] is given as :

$$E(m) = \sum_{i=1}^{N} \left[y^{(i)}(m) - a_m^T x^{(i)} - b(m) \right]^2$$

(2.19)

where a_m^T once again corresponds to the *m*th row of **A** and $\mathbf{y^{(i)}(m)}$ and **b(m)** are the *m*th components of $\mathbf{y^{(i)}}$ and **b**. The minimization of equation *2.19* with respect to $\boldsymbol{a_m}$ and **b(m)** [11] is shown as follows:

$$E(m) = \sum_{i=1}^{N} \left\{ y^{(i)}(m) - a_m^T x^{(i)} - b(m) \right\} \left\{ y^{(i)}(m) - x^{(i)^T} a_m - b(m) \right\}$$

$$\begin{aligned} E(m) = & \sum_{i=1}^{N} \left\{ y^{(i)}(m) \right\}^2 \\ & - 2a_m^T \sum y^{(i)} x^{(i)} \\ & - 2b(m) \sum y^{(i)}(m) \\ & + a_m^T \sum x^{(i)} x^{(i)^T} a_m + 2b(m) a_m^T \sum x^{(i)} + \sum b^2(m) \end{aligned}$$

$$\frac{\partial E(m)}{\partial a_m} = -2\sum y^{(i)}(m)x^{(i)} + 2\sum x^{(i)}\,x^{(i)^T} a_m + 2b(m)\sum x^{(i)} = 0$$

$$\frac{\partial E(m)}{\partial b(m)} = -2\sum y^{(i)}(m) + 2a_m^T \sum x^{(i)} + 2\sum b(m)$$

(2.20)

This results in the system of equations given as:

$$\begin{bmatrix} \sum_{i=1}^{N} x^{(i)}\,x^{(i)^T} & \sum_{i=1}^{N} x^{(i)} \\ \sum_{i=1}^{N} x^{(i)^T} & N \end{bmatrix} \begin{bmatrix} a_m \\ b(m) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y^{(i)}(m)x^{(i)} \\ \sum_{i=1}^{N} y^{(i)}(m) \end{bmatrix}$$

(2.21)

So the function *E(m)* is minimized for $m = 1$ to p. Therefore there are m different systems of equations of dimension $(p + 1)$ are solved. It is noted that since the left-hand matrix of equation 2.*21* only needs to be calculated once because it is independent of m [11]. The affine transform allows for the compensation of scaling, translation, and rotation of the feature vectors which is caused by multiple types of distortion in the speech signal and generally includes the cases of speech coding distortion, additive noise distortion and communication channel distortion.

2.4.2 McCree method. A method of signal enhancement that we have referred to as the McCree method is implemented as laid out in [13]. The first step is to perform an LP analysis of the decoded speech. The second step is to pass the decoded speech through the nonrecursive filter *A(z)*. The final step is to perform LP synthesis filtering with the transmitted LPC of the input speech to the coder in order to restore the correct spectral envelope [13].

2.4.3 Fusion strategies. Fusion strategies are implemented in order to augment system performance. Different fusion methods are utilized for the SI and SV systems namely feature level fusion and score level fusion respectively. A description of these methods is separated based on the speaker recognition system.

2.4.3.1 SI system fusion. The fusion methods for the speaker identification system are feature based. A decision level fusion strategy is implemented. The decision of a given feature is its greatest log-likelihood score. The index of that score represents the corresponding speaker. The four features contribute one speaker decision for every speech utterance. The speaker that received the most votes out of the four features would become the new speaker decision for a given test utterance in decision level fusion [11].

The second fusion method for the SI system is the use of Borda count. The Borda count method allows for the log-likelihood scores for every speaker for a given test utterance to be considered. The scores are ranked from lowest to highest for individually for each feature for every test utterance and are given a new voting total based on where the corresponding score ranks [11]. The highest voting total among all the features considered will then become the new speaker decision.

2.4.3.2 SV system fusion. Score level fusion is implemented for the SV system using the log likelihood scores from the features. Since the scores vary greatly in numeric value it is necessary to normalize the scores before the fusion processes are implemented. This is accomplished by mapping all of the scores for a single feature on the interval of 0 to 1. Where the highest score is 1 and the lowest score is 0. Each feature is normalized individually. These new normalized scores are used in the three score fusion techniques

implemented for the SV system [15]. The three score fusion techniques in the SV system are sum, product, and maximum.

Sum fusion is computed by directly summing the scores the individual features which results in a final score S_{final}. This is shown in the following equation.

$$S_{final} = \sum_{i=1}^{n} S_i$$

(2.22)

where S_i is all of the normalized feature scores and $n = 4$ since there are four features [15].

Product fusion is computed by multiplying the scores of the individual features [15] depicted in the following equation.

$$S_{final} \prod_{i=1}^{n} S_i$$

(2.23)

where S_i is all of the normalized feature scores and $n = 4$.

Max fusion is computed by taking the maximum score from all features as the final score [15].

$$S_{final} = \max(S_1, S_2, \dots, S_n)$$

(2.24)

where $n = 4$.

2.5 Statistical Analysis

A statistical analysis is required in order to prove the statistical significance of the results obtained from the speaker recognition experiments. A t test and two-way analysis of variance (ANOVA) followed by a multiple pairwise comparison are considered. All of the statistical methods described make use of a 95% confidence interval.

A two-sample t-test with unequal variances is performed to determine if the performance on clean speech is significantly better than the methods and techniques proposed in this thesis.

A two-way ANOVA allows for the analysis of two factors (feature and method) in which we can determine if there is a statistical difference among levels in the first factor, among levels in the second factor, and to see if there is an interaction effect between the two factors [16].

A multiple comparison procedure is implemented based on Tukey's procedure which enables comparison among all the group means which in turn allows us to choose the optimal combination of factors with statistical certainty [16].

Chapter 3

Approach and Methodology

Chapter 3 will detail the design approach and methodology of both speaker recognition systems. A description of the dataset partitioning, training procedure, and feature extraction process will be provided. A description of shared experimental testing protocol will be described. The experimental protocol for the SI and SV systems will be provided in full. The chapter will also discuss the SI and SV performance measures and fusion strategies. A discussion of the variation of system parameters will be included. The generation of multiple experimental trials and the application of statistical techniques to determine statistical significance will be discussed.

3.1 Dataset Initialization

The TIMIT database is used for both training and testing. All of the speech utterances for training and testing that are used from the TIMIT database are down sampled to 8 KHz prior to use in the speaker recognition systems. First, a separate partition of 168 unique speakers each having 10 speech utterances of the TIMIT database is set aside for training of the UBM. All 10 speech utterances from these 168 speakers are used in the training of the UBM. These 168 speakers will represent an alternative hypothesis or imposter model. The UBM is basically one large GMM. Another separate partition of 90 unique speakers of the TIMIT database also consisting of 10 speech utterances is used for the enrollment of the speaker recognition systems. These 90 speakers have their 10 respective utterances separated with 8 used for training and 2 used

for testing. There will be one GMM model for each speaker for a total of 90 GMMs. This set of 90 GMMs are different for each feature.

3.2 Training Phase

Consider a clean speech utterance from the TIMIT database as input. A total of 8 speech utterances are used to train a single GMM speaker model. This process is repeated once for each of the 90 speakers in the training phase.

3.2.1 Feature extraction. A speech utterance is divided into frames of 30 ms duration with a 20 ms overlap. Linear predictive analysis is performed in that the autocorrelation method is used to get a 12^{th} order LP polynomial. The LP coefficients are then converted into a 12 dimensional CEP, ACW and PST feature vector. The MFCC feature is computed using a DFT followed by a cepstral analysis using a DCT. For each of the four features, a 12 dimensional first derivative (delta) feature and second derivative (delta delta) feature is computed in each frame using a frame span of 5 (frame plus look ahead/behind 2). An energy thresholding process is performed on these 36 dimensional feature vectors where the sections of the utterance with low energy are removed [21]. Segments of silence must be removed so that only meaningful speech information contributes to the speech features. This energy thresholding process is performed on each utterance such that frames of relatively high energy corresponding to speech are identified and used to compute the feature vectors.

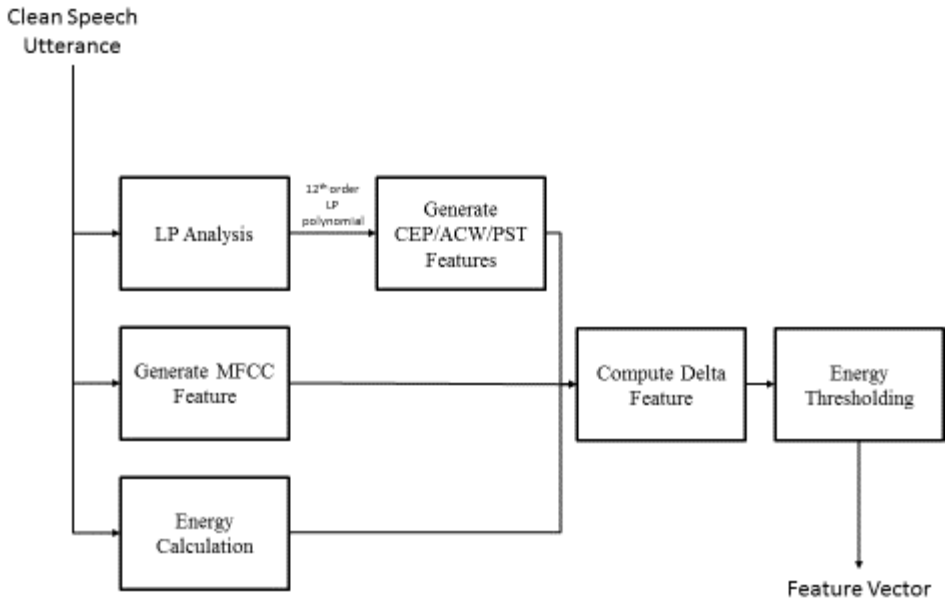

Figure 3.1. Feature extraction process

3.2.2 UBM computation. A UBM is randomly seeded by using five iterations of the k-means algorithm to initialize the parameters of an **M** mixture GMM speaker model with a diagonal covariance matrix [12]. A total of 10 iterations of the EM algorithm are performed which results in a refined GMM model. A UBM is calculated for each feature for the selected number of mixtures.

3.2.3 Individual GMM computation. The individual speaker models are obtained by MAP estimation of the UBM parameters. The calculation of these parameters are based on the designated option which is either to use all parameters (weights, means, and covariances) or to just use means. As stated previously, eight utterances are used in the training phase to obtain the feature vectors and perform the MAP adaptation.

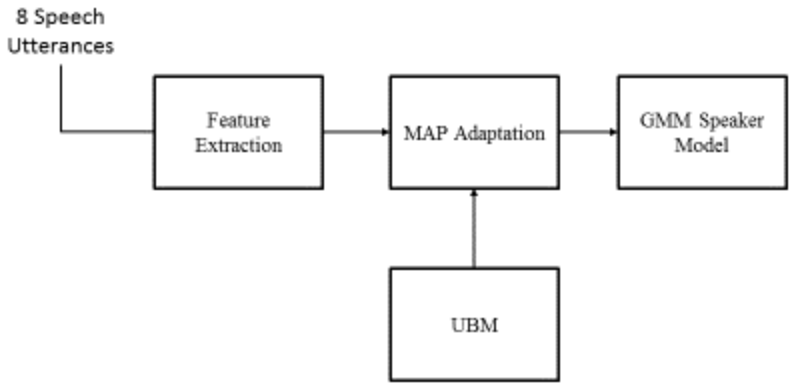

Figure 3.2. Training of a GMM speaker model

3.3 Testing Phase

Consider a clean speech utterance from the TIMIT database as input. There are two designated utterances for testing of the speaker recognition systems for each of the 90 speakers. The rotation of these utterances is described later in this chapter.

The feature extraction process is the same for training and testing for both the speaker identification system and speaker verification system with a few exceptions that allow for coder and enhancement selections. First, the test utterance is encoded with the desired speech coder (G729 8 kbit/s, G723.1 5.3 kbit/s, or GSM AMR 12.2 kbit/s). The method of enhancement is then chosen (no enhancement, McCree method, affine transform, both McCree and affine). Note that the affine transformation applied after the feature extraction is performed as shown in the following figure.

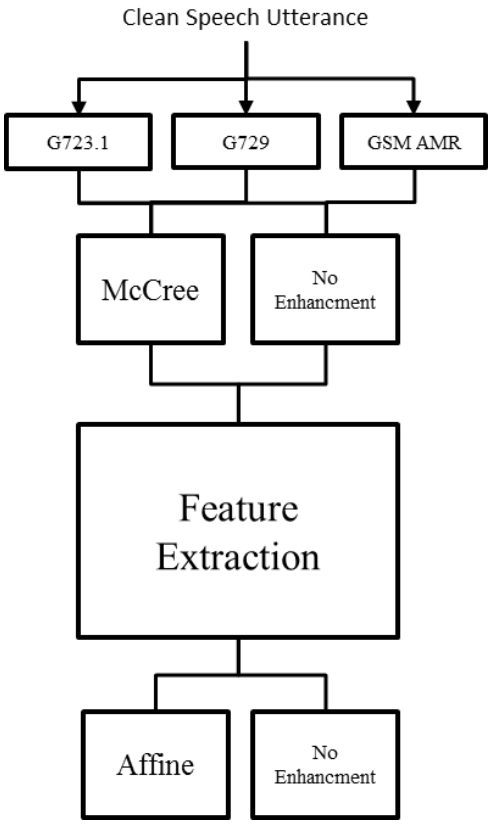

Figure 3.3. Testing phase enhancement diagram

3.3.1 Enhancement methods. An established signal enhancement method as well as a novel feature enhancement method are investigated.

3.3.1.1 McCree method. The test utterances for each coder type have the McCree method of signal enhancement applied prior to the start of the testing phase. The test utterance for the desired coder where the McCree method is applied is used when the McCree method is selected.

3.3.1.2 Affine transform. The affine transform parameters are calculated from the first 5 training utterances. These utterances are reserved for the affine transform and are not affected by the rotation of the testing data which will be described later in this

chapter. The first and second derivative information are not used in the calculation of the affine transform. The affine transform is computed prior to the testing phase. There is a unique affine transform for each of the four features for all three coders. In addition, there is also a unique affine transform if the McCree method is selected for every feature and coder combination.

3.3.1.3 McCree method and affine transform. A combination of enhancement methods is performed. The test utterances with the McCree method applied are used with their corresponding affine transform based on feature and coder selection.

3.3.2 Speaker recognition system experimental protocol. The testing phase experimental protocol for the speaker identification system and speaker verification system that is not shared is described in this section in detail.

3.3.2.1 Speaker identification system. The decision logic for the SI system is implemented after the feature extraction process is complete and all of selected enhancement methods are applied. The SI system attempts to solve a 1:*M* speaker problem where $M = 90$. The objective of the SI system is to determine which speaker's GMM model out of the 90 total speaker models is closest to the input test utterance's feature vectors.

There are $M = 90$ speakers for which speaker *i* is represented by GMM λ_i. M^* is the identified speaker and is chosen to maximize the a posteriori log-probability [11] as shown in the following equation.

$$M^* = \arg \max_{1 \le j \le M} \sum_{i=1}^{q} \log p(x_i|\lambda_j)$$

(3.1)

where $p(x_i|\lambda_j)$ is computed as given in equation 2.15. If the identified speaker matches the actual speaker of the test utterance in question, it is recorded as a correct identification.

3.3.2.1.1 Speaker identification performance measure. The performance of the speaker identification system is measured using the identification success rate (ISR). The ISR is represented as the total number of correct identifications divided by the total number of test trials. In a single experimental procedure, there are 90 speakers which have two test utterances each which totals for 180 test cases. This process is repeated for all possible variations of system parameters in which the ISR is calculated independently for each parameter variation.

3.3.2.2 Speaker verification system. The decision logic for the SV system is also implemented after the feature extraction process is complete and all of selected enhancement methods are applied. The SV system attempts to solve a 1:1 speaker problem where we determine if the test utterance's feature vectors are a close enough match to the claimed identity's speaker model based on a threshold to either accept or reject the claimed identity.

Let the claimed identity of a speaker be *k*. The posteriori log-probability as in equation 3.1 is computed for the speaker model λ_k and for the UBM model. The SV score is calculated by subtracting the speaker model score λ_k by the UBM score. For

each feature and for each coder there will be 180 genuine or true attempts where the test utterance is actually the claimed identity and there will be 16,020 imposter attempts where the test utterance is not actually the claimed identity. Table 3.1 details the true and imposter attempts below.

Table 3.1

True/imposter attempt breakdown

Type	**True**	**Imposter**
Total Number of Attempts	180	16,020
Explanation	(2)(90) 2 utterances for each speaker	(2)(90)(89) 2 utterances for each of the 90 speakers 89 times each attempt

3.3.2.2.1 Speaker verification performance measure. The SV score is compared to a threshold to either accept or reject the claimed identity. The false accept rate (FAR) and false reject rate (FRR) are adjusted based on the threshold chosen which in turn yields a receiver operating characteristic (ROC) from which the equal error rate is the performance measure. The EER being the point on the ROC in which the FAR equals the FRR. Once again this testing process is repeated for all possible variations of system parameters in which the EER is calculated independently for each parameter variation.

3.3.3 Variation of parameters. The four methods under investigation in this thesis are to perform no enhancement, to perform signal enhancement (McCree method), to perform feature enhancement (affine transform), or to perform both enhancements (McCree method and affine transform). The data set was exhaustively tested for each of our four methods for both the SI and SV systems by varying the following parameters.

The type of speech coder is varied which include the G723.1 speech coder (5.3 kbps), the G729 speech coder (8 kbps), and the GSM AMR speech coder (12.2 kbps selection).

The number of Gaussian mixtures used for the speaker models was varied from 16 to 2048 in powers of two (16, 32, 64, 128, 256, 512, 1024, 2048). The GMM speaker model is tested with a UBM with the corresponding number of mixtures. So a GMM model tested on 16 mixtures is tested with a UBM with 16 mixtures.

For MAP estimation, there are two options. One is to use all parameters (weights means and covariances) and the other option is to just adapt the means only.

Four features are examined, namely, CEP, ACW, PST, and MFCC.

3.3.4 Fusion methods. Different fusion methods were utilized for both speaker recognition systems. A description of these methods is separated based on the speaker recognition system. Each coder and method of enhancement are considered independent for all fusion methods.

3.3.4.1 Speaker identification system fusion methods. The fusion methods for the SI system are feature based. Every combination of feature is considered in the fusion methods as described in the following table. A final selection of features to be used in the SI fusion methods will be determined experimentally.

Table 3.2

Feature fusion possibilities

Feature List	**Fusion Name**
CEP, ACW, PST, MFCC	CAPM
CEP, ACW, PST	CAP
CEP, ACW, MFCC	CAM
ACW, PST, MFCC	APM
CEP, ACW	CA
CEP, PST	CP
CEP, MFCC	CM
ACW, PST	AP
ACW, MFCC	AM
PST, MFCC	PM

3.3.4.1.1 Decision level fusion. The four features (CEP, ACW, PST, MFCC) final speaker decision are considered where the speaker with the most final decision votes become the new decision. A tie (1-1-1-1 or 2-2) is resolved by arbitrarily taking the lowest speaker number as the final decision.

3.3.4.1.2 Borda count fusion. Borda count fusion considers all of the speakers as a possible decision instead of only counting the final decision from each feature. The speakers are ranked from lowest to highest in log-likelihood score and are then assigned a new score based on their cumulative ranking amongst all the features in question. Since

all 90 speakers are eligible it is now possible for a speaker that has scored higher on a few features but not the highest to be chosen as the final decision.

3.3.4.2 Speaker verification system fusion methods. The fusion methods for the SV system are score based. The score fusion methods in this thesis are considered combinational approaches and it is necessary to perform a score normalization before fusion [15]. The scores have a great variation of values due to its logarithmic basis. In order to accurately represent the normalized scores the following equation is used to calculate a normalized score *y*.

$$y = \frac{(x - x_{min})}{x_{max} - x_{min}}$$

3.2

where x is the raw score and x_{min} and x_{max} are the minimum and maximum scores of a single feature and type of score (true or imposter). This equation is implemented for the true scores and the imposter scores separately on a feature by feature basis. Once the score normalization takes place a score fusion method can be implemented. The three methods used in this thesis are to directly add the scores (sum fusion), multiply the scores (product fusion), or to take the maximum value of the scores (maximum fusion). The scores of all four features are considered when performing score fusion.

3.4 Statistical Analysis

In order to perform a statistical analysis, multiple experiment trials are needed in order to determine if the results obtained are statistically significant. These trials are formed by rotating the testing and training utterances. A total of 10 trials are conducted

per method for each speech coder. The last 5 speech utterances for each speaker are rotated since the first 5 utterances are reserved for the calculation of the affine transform. These 10 trials will be performed on a finalized number of Gaussian mixtures as well as the MAP adaptation option that have been experimentally determined to be optimal or near optimal compared to the rest of the possible parameters. The following table breaks down how the test utterances are used for training and testing for a given speaker.

Table 3.3

Training and testing utterance convention

Trial Number	**Training Utterances**			**Testing Utterances**	
1	8	9	10	6	7
2	7	9	10	6	8
3	7	8	10	6	9
4	7	8	9	6	10
5	6	9	10	7	8
6	6	8	10	7	9
7	6	8	9	7	10
8	6	7	10	8	9
9	6	7	9	8	10
10	6	7	8	9	10

Note: Utterances 1-5 are always used in training since they are used for when calculating the affine transform

3.4.1 Two-Factor ANOVA. A two-factor or two-way analysis of variance (ANOVA) is utilized to prove statistical significance [16]. The two factors that are under investigation are feature and method. These two factors are tested independently for both the SI and SV systems and are also tested with and without the application of fusion strategies. For the purposes of the ANOVA, a fusion strategy is considered to be another

feature. So for example, decision level fusion and Borda count are considered additional features for the SI system and the score fusion methods of sum, product, and maximum are considered additional features for the SV system. The four methods investigated in this thesis are to perform no enhancement, to perform the McCree method (signal enhancement), to perform the affine transform (feature enhancement), and to perform both the McCree method and affine transform. The table below details the possible feature combinations.

Table 3.4

Features and fusion description

Speaker Recognition System	**Features without Fusion**				**Additional Features with Fusion**		
SI	CEP	ACW	PST	MFCC	Decision level		Borda count
SV	CEP	ACW	PST	MFCC	Sum	Product	Max

The three coders used (G729, G723.1, and GSM AMR 12.2) are considered to be separate distributions so that a two-way ANOVA is performed for each coder. A total of 12 two-way ANOVA's are performed to consider all possible test scenarios in order to determine the optimum feature and optimum method selection for each speech coder, speaker recognition system, and based on the inclusion or exclusion of fusion strategies. The completion of this process will show if the results obtained are statistically significant. The two-way ANOVA will show whether or not there is a statistical

difference among the features, among the methods, and also if there is an interaction effect between the feature and the method for a given distribution.

3.4.2 Multiple comparison procedure. Further analysis is required in order to identify which pairs of feature and method are significantly different from one another. This is accomplished by use of a multiple comparison test specifically using the Tukey-Kramer method [16]. Observing the difference in the pairwise comparison of group means allows for the determination of the optimum feature and optimum method selection. A confidence interval of 95% is used in the multiple comparison test.

Chapter 4

Results

This chapter will contain a comprehensive presentation of the results of the many experiments conducted in this thesis. The finalization of initial parameters and the scope of experiments performed is explored. The results of the speaker identification system and speaker verification system in terms of average identification success rate and average equal error rate respectively is detailed. Section 4.3 describes the statistical analysis of these results. This includes a multiple comparison procedure that examines both enhancement method and feature selection for both the SI and SV system for a 95% confidence interval. A two sample t-test is performed on the best approach for each coder on both speaker recognition systems and compared to the performance of a clean speech benchmark.

4.1 Initial Parameters

In preparation for multiple experiment trials it is first necessary to determine optimal initial parameters. The number of Gaussian mixtures and MAP adaptation option are examined. These initial parameters are determined experimentally. When determining initial parameters only one trial is performed instead of a total of 10 (Trial number 10 is performed). There are 64 experimental trials per feature which makes for 256 experimental trials for each coder type for a grand total of 768 preliminary trials. Optimal initial parameters can be determined experimentally through analysis of these preliminary trials. Table 4.1 depicts a detailed breakdown of the preliminary trial possibilities.

Table 4.1

Preliminary experiment variations

Testing Variables	**Amount**	**Details**
Coding Distortion	3	G723.1, G729, GSM-AMR
Features	4	CEP, ACW, PST, MFCC
Method of Enhancement	4	No Enhancement, McCree, Affine, McCree & Affine
Number of Gaussian Mixtures	8	16, 32, 64, 128, 256, 512, 1024, 2048
MAP Adaptation Option	2	Use All Parameters or Use Means only
Number of Trials	1	Trial 10 only
Total Preliminary Experiments	768	(3)(4)(4)(8)(2)(1)

The number of mixtures was varied from 16 to 2048 in powers of 2. The use of 128, 256 and 512 mixtures yielded the best comparable performance. This is depicted for the CEP feature for the SI system in figure 4.1 and the SV system in figure 4.2. This holds true for all four features. Note that a superior ISR value is greater when considering the performance of the SI system and a superior EER value is lower when considering the performance of the SV system.

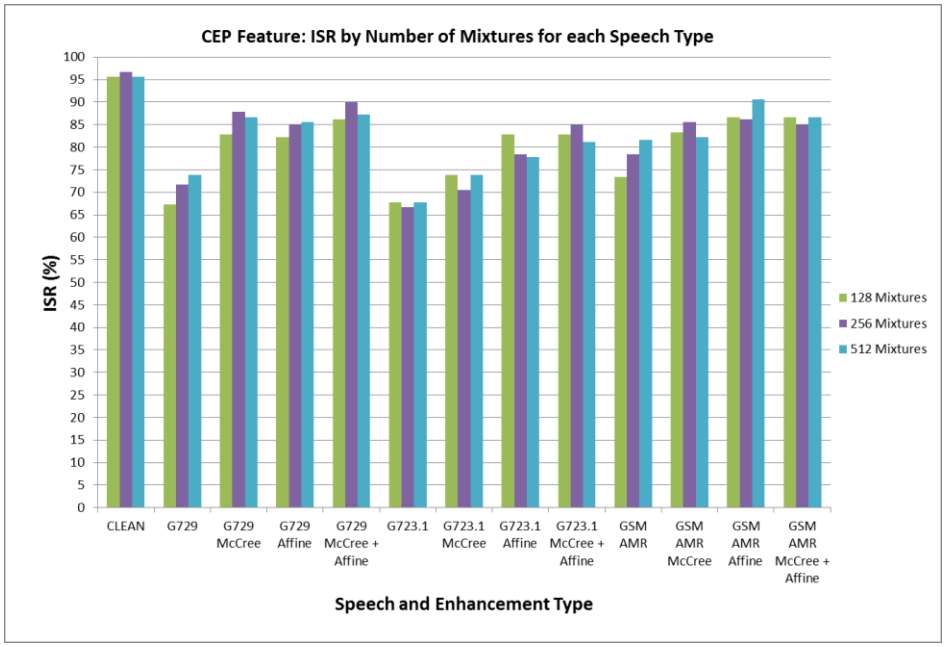

Figure 4.1. Mixture selection ISR for CEP feature. Depicted are 128, 256, and 512 mixtures for each speech type and enhancement method combination. Note that a superior or desirable ISR value is one that is greater.

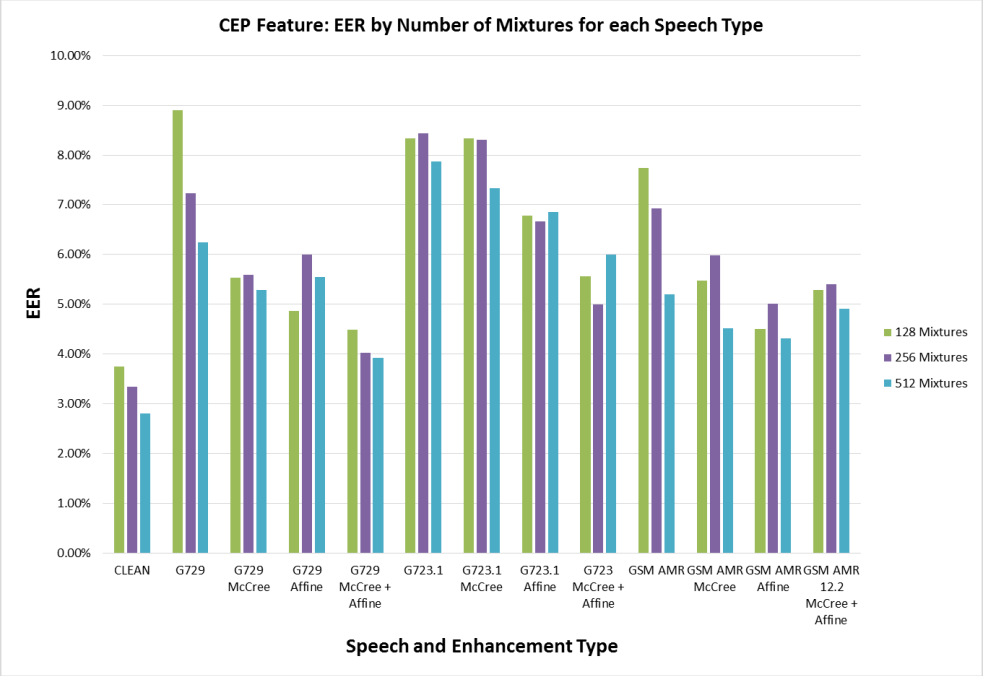

Figure 4.2. Mixture selection EER for CEP feature. Depicted are 128, 256, 512 mixtures for each speech type and enhancement method combination. Note that a superior or desirable EER value is one that is lesser.

Using more than 512 mixtures resulted in additional computational complexity and did not necessarily improve performance. The usage of a greater number of mixtures results in diminishing returns in system performance. This is supported by [12]. Therefore the number of Gaussian mixtures is set at 256. It was experimentally found that it was only necessary to use the means when performing MAP adaptation. This determination is also supported by [12]. This fact is shown graphically for the SI system in figure 4.3 and the SV system in figure 4.4. This also holds true for all four features.

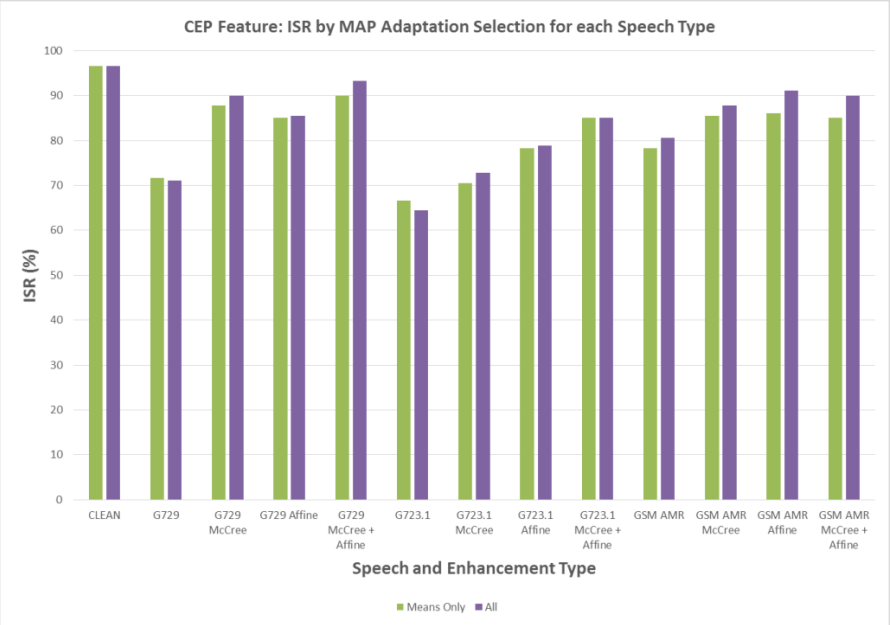

Figure 4.3. MAP adaptation selection ISR for CEP feature. Depicted is 256 mixtures for each speech type and enhancement method combination. Note that a superior or desirable ISR value is one that is greater.

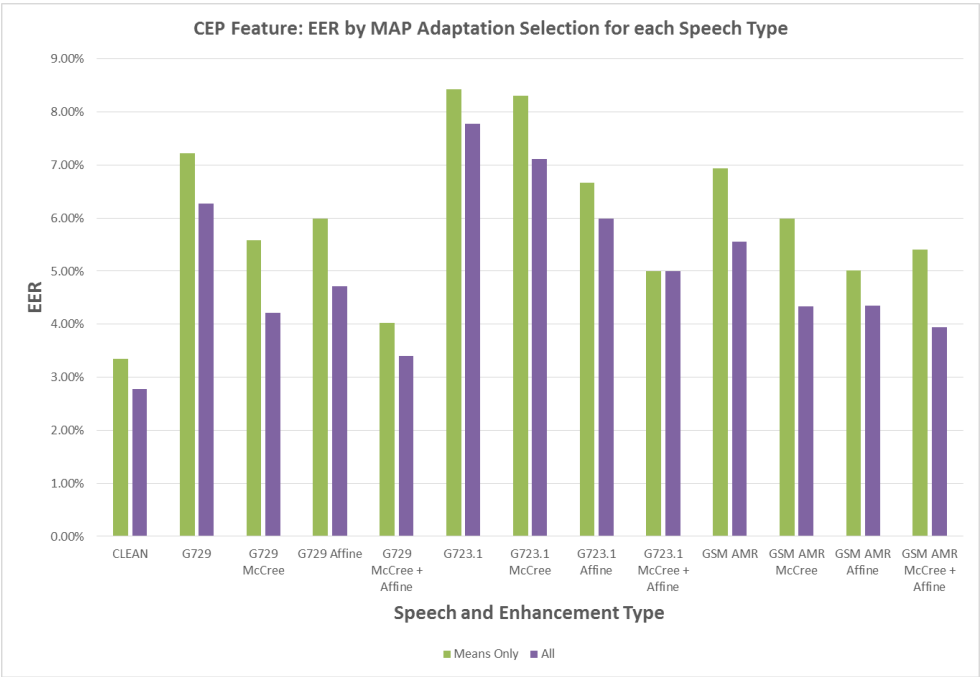

Figure 4.4. MAP adaptation selection EER for CEP feature. Depicted is 256 mixtures for each speech type and enhancement method combination. Note that a superior or desirable EER value is one that is lesser.

Once experimental parameters have been finalized the testing phase can be implemented. There are 16 experiments conducted for each coder. Each experiment is repeated 10 times by rotating the training and testing utterances as described in table 3.3. This results in 160 experiments for each coder for a total of 480 experiments. This experimental protocol is performed on the SI and SV system separately in the ways described previously in sections 3.3.2.1 and 3.3.2.2 respectively. A description of the testing possibilities are described below in table 4.2.

Table 4.2

Finalized testing variations

Testing Variables	**Amount**	**Details**
Coding Distortion	3	G723.1, G729, GSM-AMR
Features	4	CEP, ACW, PST, MFCC
Method of Enhancement	4	No Enhancement, McCree, Affine, McCree & Affine
Number of Gaussian Mixtures	1	256
MAP Adaptation Option	1	Use Means only
Number of Trials	10	Trials 1 through 10
Total Preliminary Experiments	480	$(3)(4)(4)(1)(1)(10)$

4.2 Speaker Recognition System Results

The following section details the results from the experiments conducted for the SI system and SV system in terms of average ISR and EER respectively. Further analysis of these results is conducted in Section 4.3 in form of a two-way ANOVA followed by a multiple comparison test and a two sample t-test.

4.2.1 Speaker identification system results. Table 4.3 contains the average ISR for a given condition and feature over 10 trials. A test on clean speech (no coder distortion added) is performed for comparison. Each coder is tested for all four features (CEP, ACW, PST, MFCC) and for all methods of enhancement (no enhancement, McCree signal enhancement, affine transform feature enhancement, McCree signal enhancement combined with affine feature enhancement).The two feature fusion methods, decision level and Borda count, consider all four features when determining the fused ISR and also represent an average over 10 trials. The feature fusion methods add an additional 240 experiments to the overall SI system experiment total (80 for each coder).

Table 4.3

ISR for all testing conditions

Condition	CEP	ACW	PST	MFCC	Decision Fusion	Borda Count
Clean	93.1	93.2	92.1	95.4	95.2	95.2
G723.1	64.6	62.5	65.3	79.3	69.0	72.3
G723.1 McCree	70.4	67.6	71.2	83.0	75.3	77.3
G723.1 Affine	77.7	74.2	77.7	86.3	83.4	84.3
G723.1 McCree + Affine	82.8	78.9	80.7	85.8	86.5	87.9
G729	65.7	61.4	64.6	78.5	69.9	70.2
G729 McCree	85.0	83.5	83.6	91.1	88.3	89.3
G729 Affine	84.3	80.9	82.1	89.3	87.8	88.9
G729 McCree + Affine	86.8	85.5	86.7	90.3	90.2	91.1
GSM-AMR	75.9	73.8	75.3	78.9	78.9	76.3
GSM-AMR McCree	86.1	83.7	84.2	84.2	87.7	84.4
GSM-AMR Affine	86.5	86.6	86.2	84.0	89.8	85.3
GSM-AMR McCree + Affine	85.3	84.2	83.8	83.6	88.2	83.8

Note: Each ISR is in the form of an average percentage over 10 trials for a given condition. McCree + Affine refers to using the combination of both enhancement methods.

4.2.2 Speaker verification system results. Table 4.4 contains the average EER for a given condition and feature over 10 trials. A test on clean speech (no coder distortion added) is also performed for comparison. Once again, each coder is tested for all four features and for all methods of enhancement. There are three score fusion methods performed, sum fusion, product fusion, and maximum fusion. These score fusion methods consider all four features when determining the fused EER and also represent an average over 10 trials. The score fusion methods add an additional 360 experiments to the overall SV system experiment total (120 for each coder).

Table 4.4

EER for all testing conditions

Condition	**CEP**	**ACW**	**PST**	**MFCC**	**Sum Fusion**	**Prod Fusion**	**Max Fusion**
Clean	3.61	3.35	3.39	3.13	2.78	2.77	3.40
G723.1	8.43	8.79	8.87	5.98	6.48	6.65	6.62
G723.1 McCree	7.87	8.09	7.67	5.43	5.75	5.88	6.01
G723.1 Affine	5.75	6.61	6.22	4.59	4.60	4.56	5.27
G723.1 McCree + Affine	4.95	5.73	5.51	4.29	4.10	4.13	4.74
G729	8.11	8.59	8.44	6.69	6.44	6.57	6.63
G729 McCree	4.82	4.85	4.80	3.90	3.74	3.67	4.12
G729 Affine	5.29	4.85	4.79	4.07	3.79	3.77	4.38
G729 McCree + Affine	4.05	4.04	3.93	3.51	3.13	3.19	3.77
GSM-AMR	6.63	6.18	6.25	4.61	4.90	4.90	5.77
GSM-AMR McCree	5.38	4.86	4.94	3.29	3.51	3.47	4.55
GSM-AMR Affine	4.65	4.58	4.55	3.44	3.26	3.24	4.37
GSM-AMR McCree + Affine	5.34	4.96	4.96	3.39	3.66	3.58	4.89

Note: Each EER is in the form of an average percentage over 10 trials for a given condition. McCree + Affine refers to using the combination of both enhancement methods.

4.3 Statistical Analysis of Results

A discussion of the comparison among the methods and features individually is given. Considering the interaction between the methods and features, the best approaches are also mentioned.

4.3.1 SI system G723.1. Figure 4.5 and 4.6 show the 95% confidence interval for the methods and features respectively. It is clear that combining the McCree technique and the affine transform is the best method. The features (includes decision level fusion and Borda count) are similarly compared and the best feature is the MFCC. Due to the interaction of the feature and method, the best performance (average ISR of 87.9%) is obtained using the McCree technique and the affine transform in conjunction with Borda count fusion.

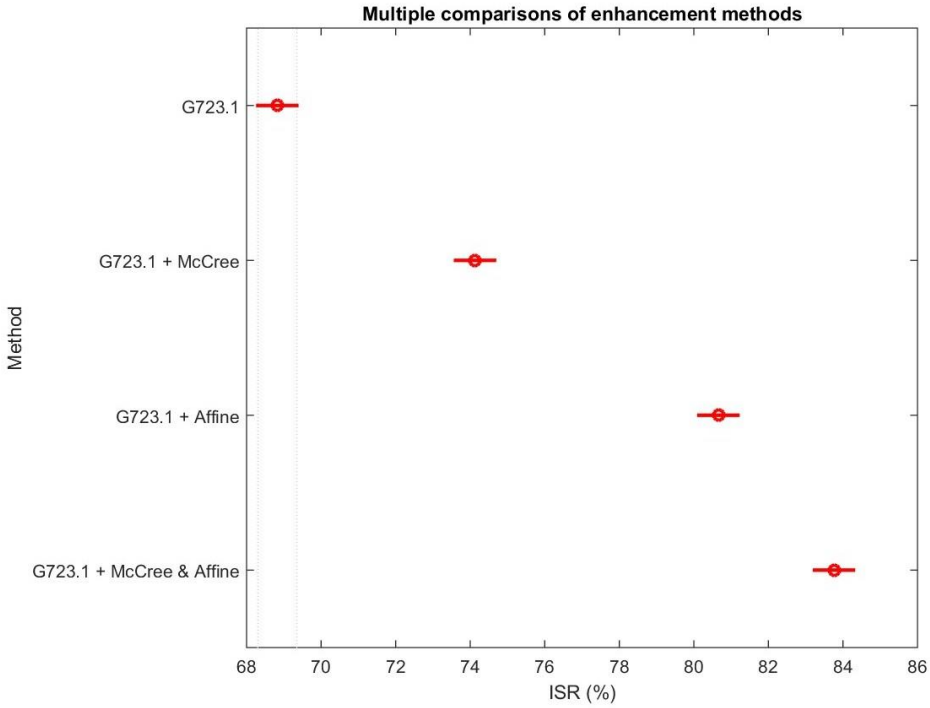

Figure 4.5. SI comparison of the methods (G723.1)

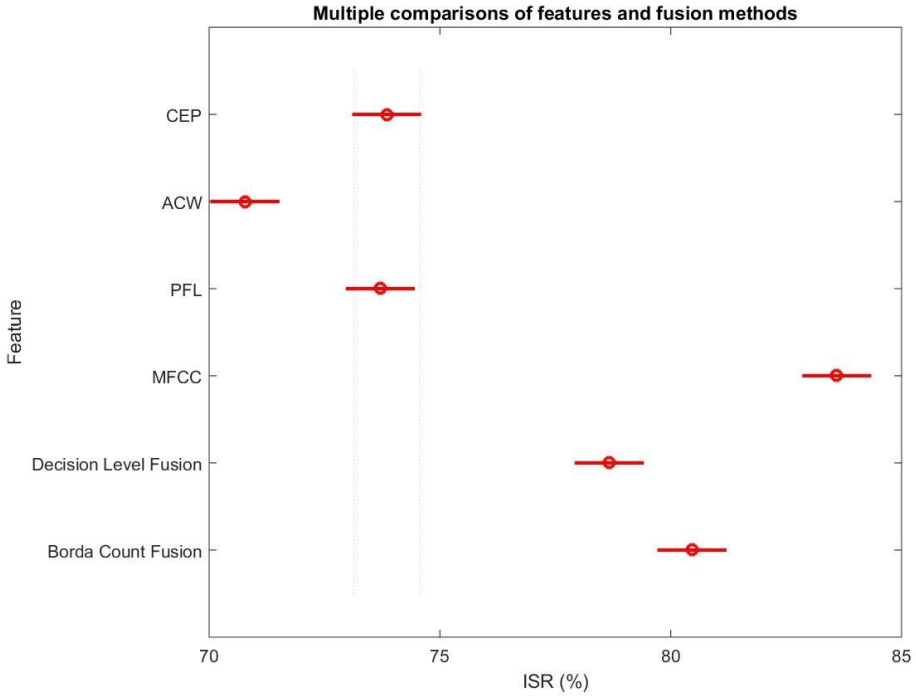

Figure 4.6. SI comparison of the features (G723.1)

4.3.2 SI system G729. Figures 4.7 and 4.8 show the 95% for the methods and features respectively. As in the case of G.723.1 the best method is to combine the McCree technique and the affine transform and the best feature is the MFCC. Due to the interaction of the feature and method, the best performance (average ISR of 91.1%) is obtained using either the McCree technique and the affine transform in conjunction with Borda count fusion or the McCree technique with the MFCC feature.

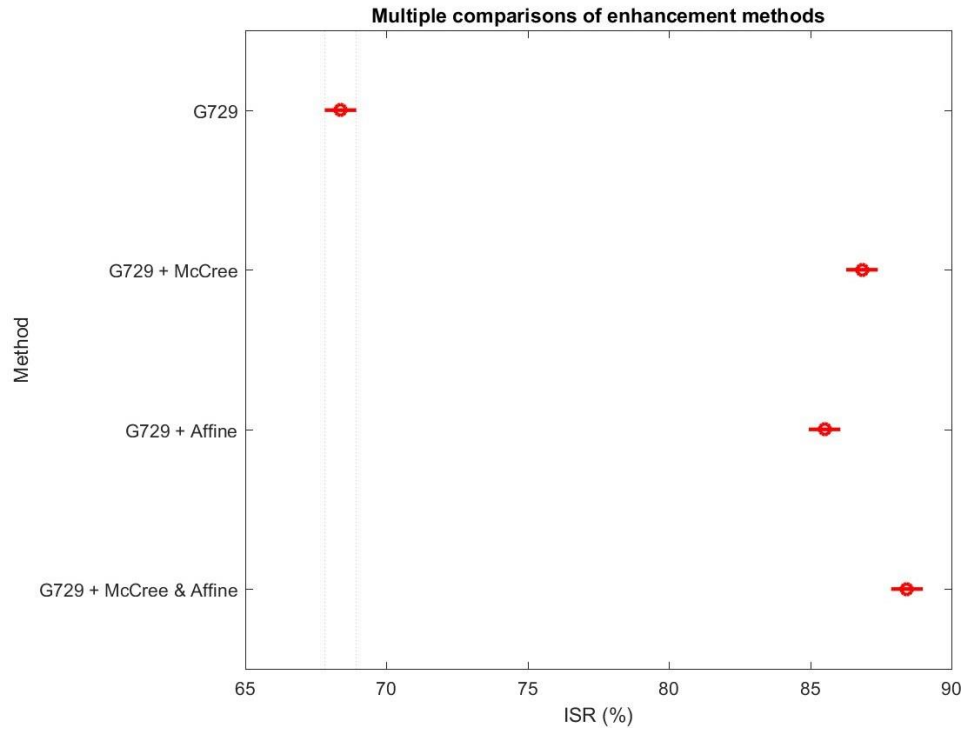

Figure 4.7. SI comparison of the methods (G729)

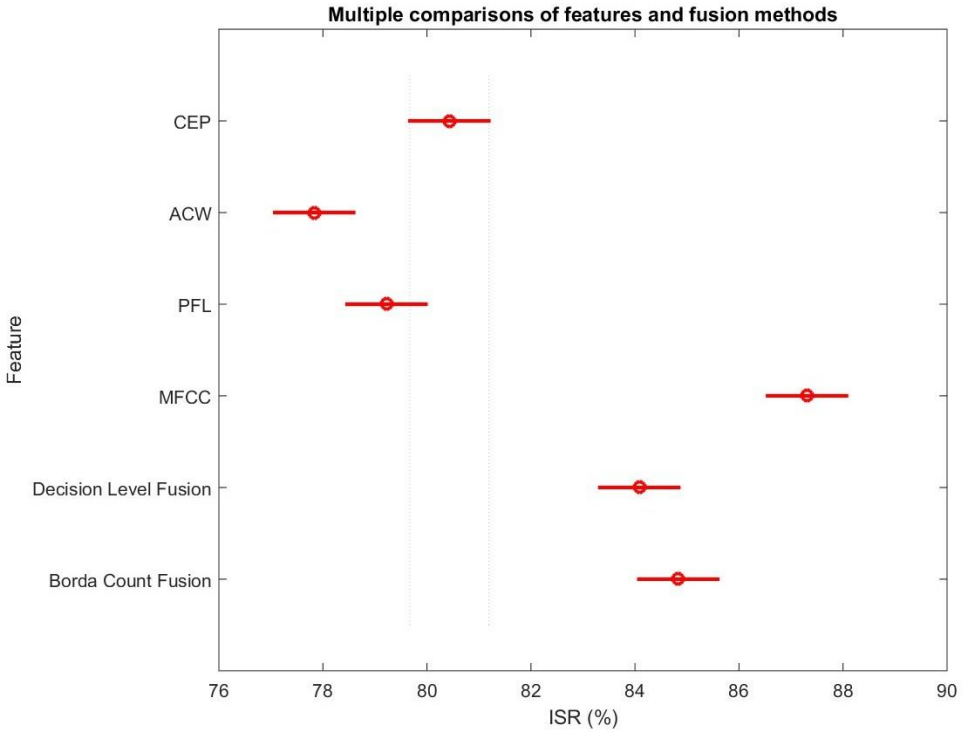

Figure 4.8. SI comparison of the features (G729)

4.3.3 SI system GSM-AMR. Figure 4.9 and 4.10 show the results. The best method is using only the affine transform. The best feature is the use of decision level fusion. It is the same two approaches that interact the best achieving an average ISR of 89.8%

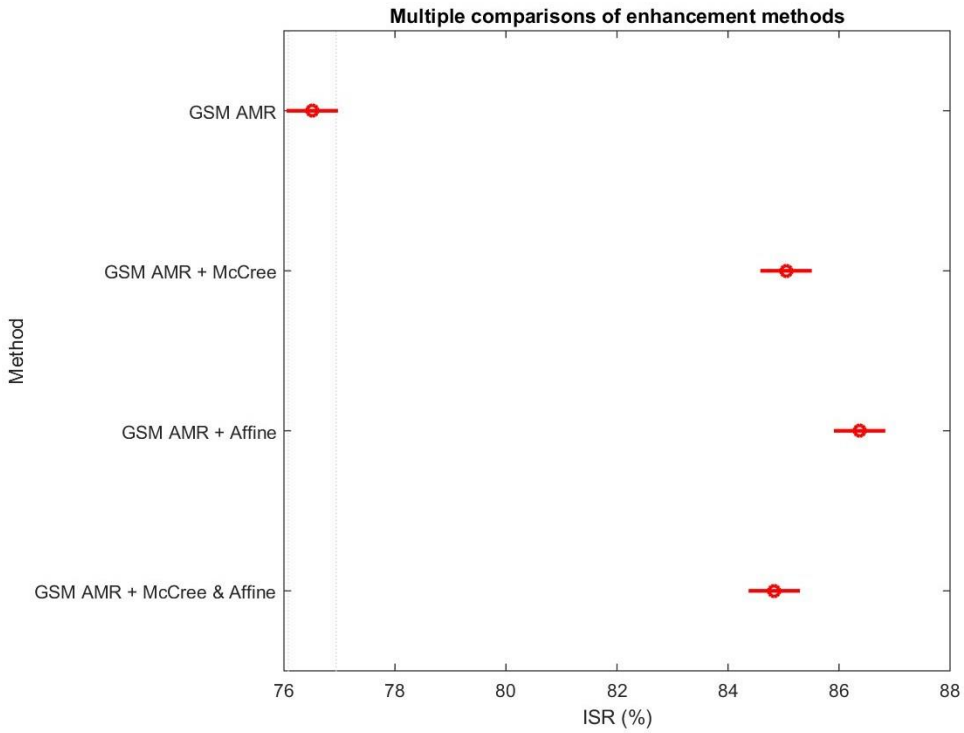

Figure 4.9. SI comparison of the methods (GSM-AMR)

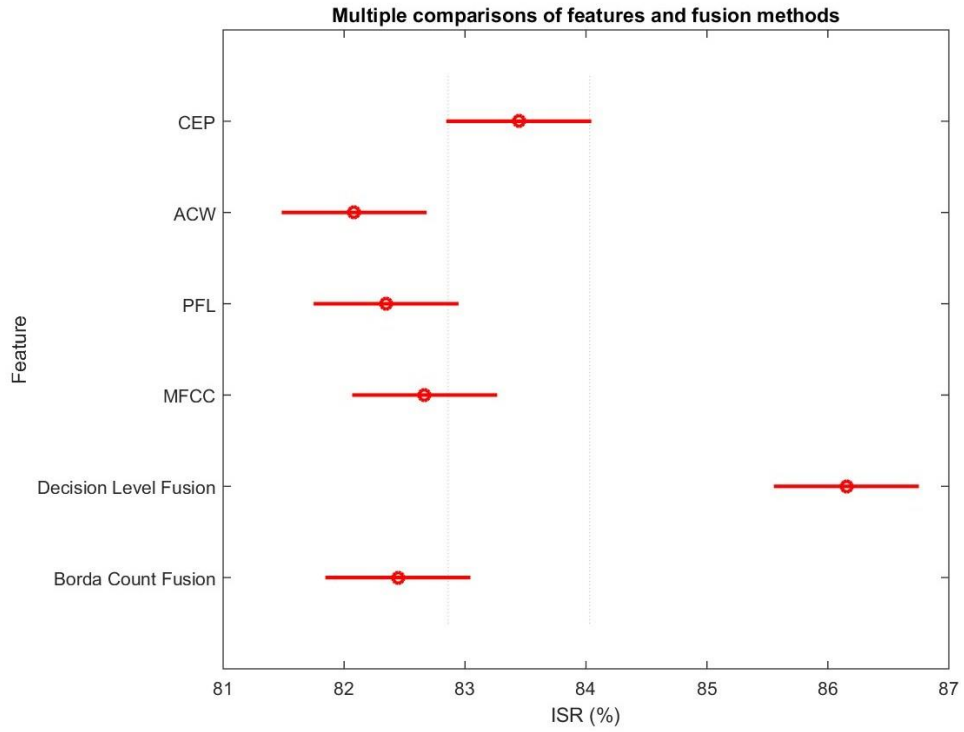

Figure 4.10. SI comparison of the features (GSM-AMR)

4.3.4 SV system G723.1. Figure 4.11 and 4.12 show the 95% confidence interval for the methods and features respectively. It is clear that combining the McCree technique and the affine transform is the best method. The features (includes sum, product and maximum score fusion) are similarly compared. Although the best feature is the MFCC, its 95% confidence interval overlaps with that of sum and product fusion. Due to the interaction of the feature and method, the best performance (average EER of 4.1%) is obtained using the McCree technique and the affine transform in conjunction with sum fusion. Using product fusion is statistically comparable and leads to only a slightly higher average EER of 4.13%.

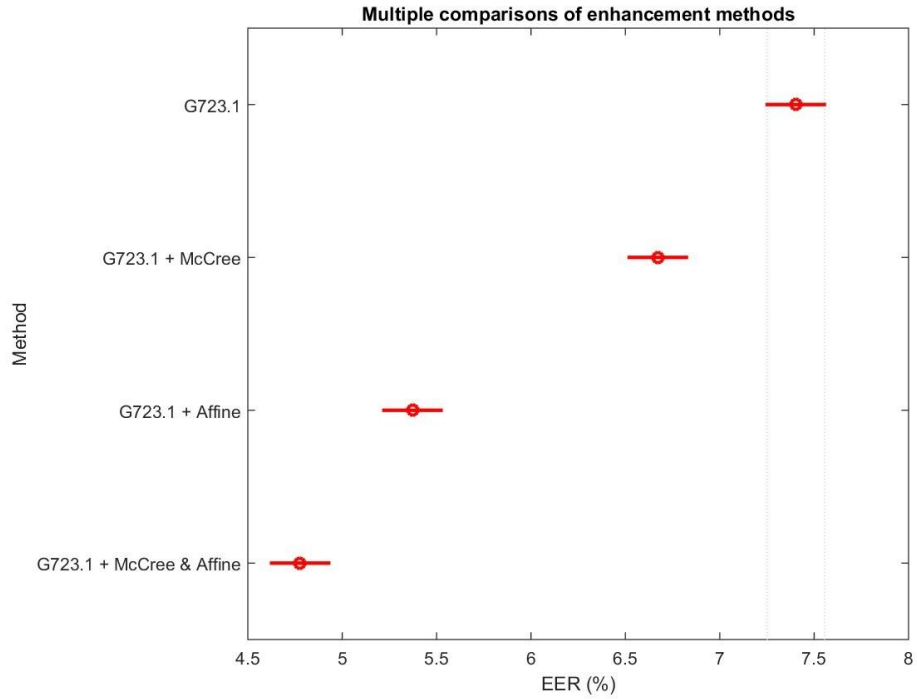

Figure 4.11. SV comparison of the methods (G723.1)

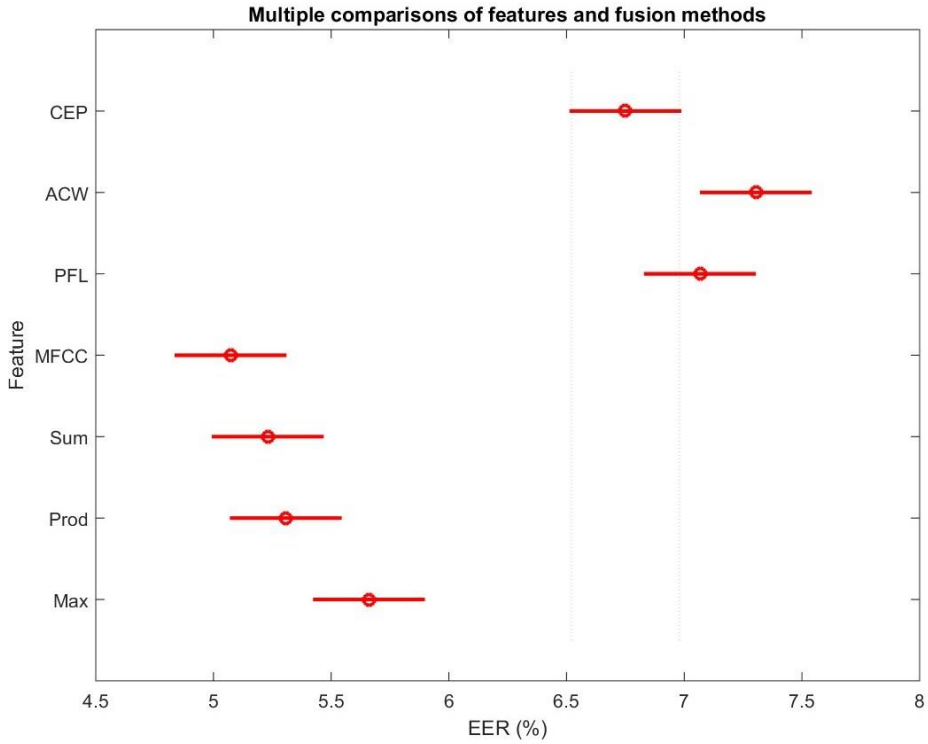

Figure 4.12. SV comparison of the features (G723.1)

4.3.5 SV system G729. Figures 4.13 and 4.14 show the 95% confidence interval for the methods and features respectively. The best method is to combine the McCree technique and the affine transform. Although sum fusion is the best feature its 95% confidence interval has considerable overlap with the product fusion and partial overlap with the MFCC. Due to the interaction of the feature and method, the best performance (average EER of 3.13%) is obtained using the McCree technique and the affine transform in conjunction with sum fusion.

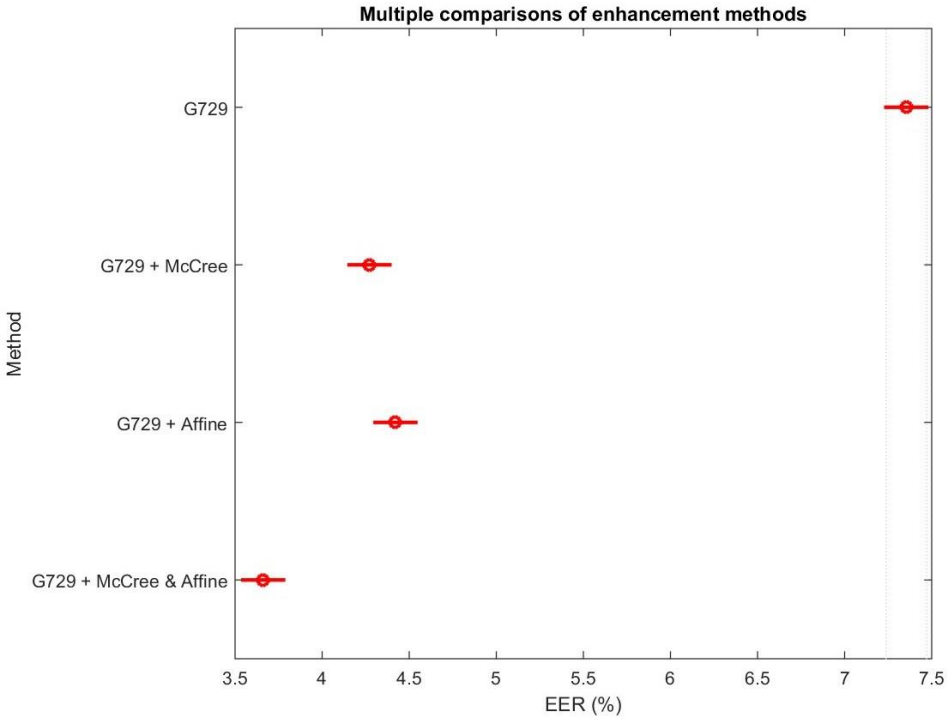

Figure 4.13. SV comparison of the methods (G729)

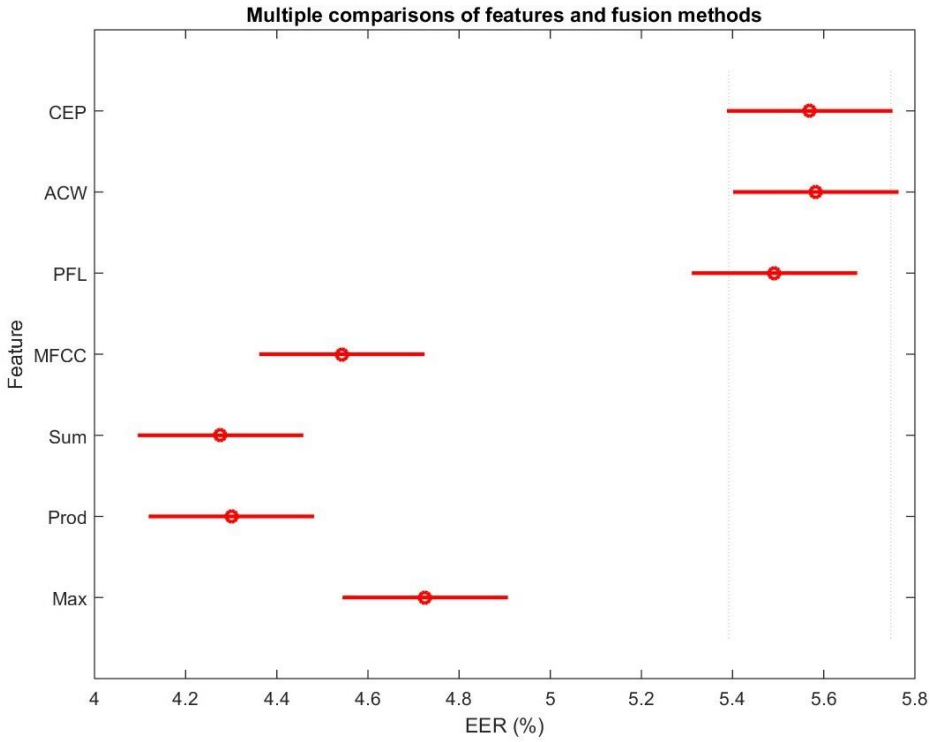

Figure 4.14. SV comparison of the features (G729)

4.3.6 SV system GSM-AMR. Figure 4.15 and 4.16 show the results. The best method is using only the affine transform. The best features are the MFCC, sum fusion and product fusion. Due to interaction, the three best approaches are MFCC with McCree (3.29%), sum fusion with affine (3.26%) and product fusion with affine (3.24%). All three are statistically indistinguishable.

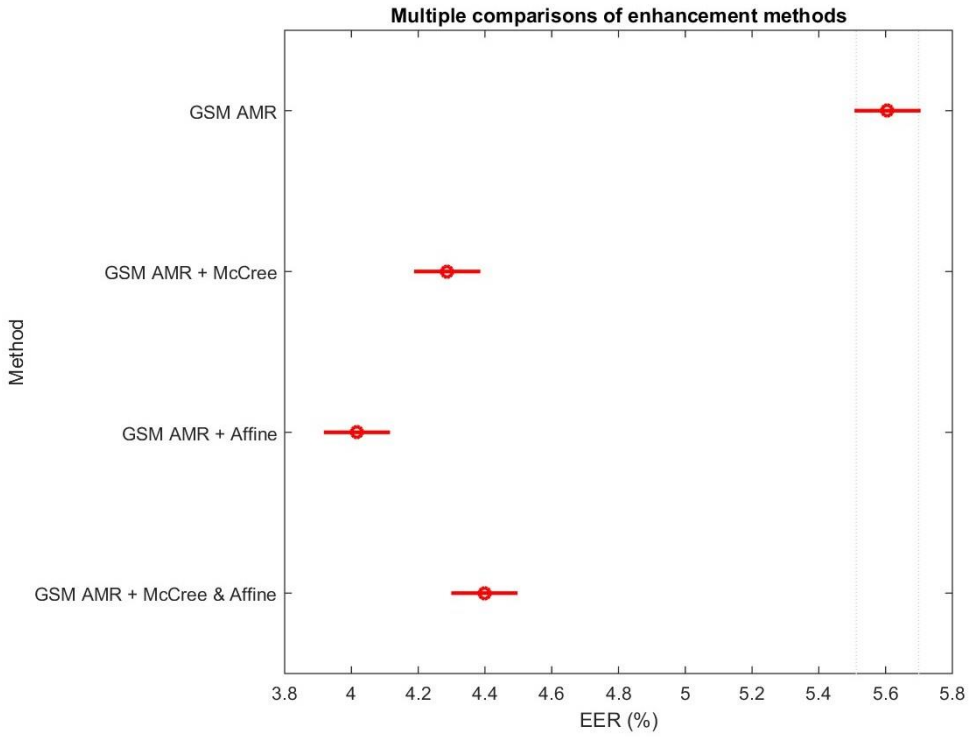

Figure 4.15. SV comparison of the methods (GSM-AMR)

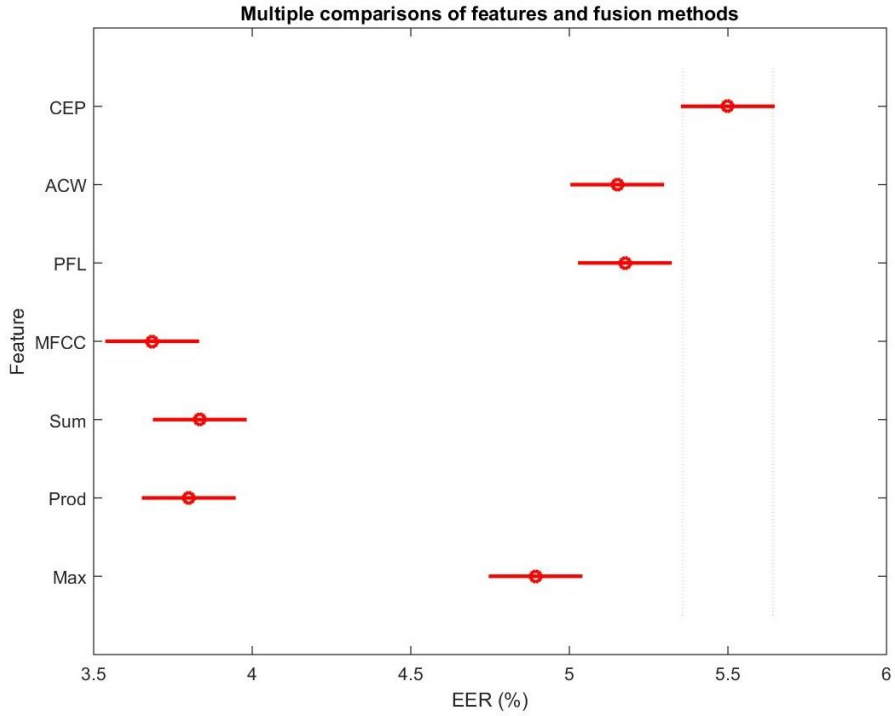

Figure 4.16. SV comparison of the features (GSM-AMR)

The following table lists the optimal feature and method selections for each coder and speaker recognition system based on the above results.

Table 4.5

Optimal selection for each system and coder grouping

Coder and System	**G723.1 SI**	**G729 SI**	**GSM-AMR SI**	**G723.1 SV**	**G729 SV**	**GSM-AMR SV**
Optimum Feature	MFCC	MFCC	Decision Level	MFCC, Sum, Product	MFCC, Sum, Product	MFCC, Sum, Product
Optimum Method	McCree + Affine	McCree + Affine	Affine	McCree + Affine	McCree + Affine	Affine

Note: The optimum feature is statistically similar when more than one feature is selected. McCree + Affine refers to using the combination of both enhancement methods.

4.3.7 Comparison with testing on clean speech. In the case of testing on clean speech, neither signal nor feature enhancement is necessary. Also, there is no statistical difference among the features and fusion methods for both SI and SV systems. The purpose is to compare the performance of the best approaches for each speech coder with the performance on clean speech. Table 4.6 gives the average ISR comparisons for the SI case. There are two approaches that achieve the best average ISR for the G.729 coder. The MFCC feature is selected as the benchmark for clean speech as it achieves the highest average ISR. The best approach for each coder is individually compared to the test case of clean speech only. Therefore, a two sample statistical t-test with a 5% significance level and unequal variances is performed to determine if the performance on

clean speech is significantly better than the technique used for each coder. The test is based on the 10 trials that are performed for a given experiment.

Table 4.6 also gives the obtained p-values. Although the methods have mitigated the train/test mismatch and led to a substantial performance improvement, the low p-values indicate that the ISR values are not statistically comparable to that of clean speech. Table 4.7 gives the average EER comparisons for the SV case. Product fusion is selected as the benchmark for clean speech as it achieves the lowest average EER. Again, the best approach for each coder is individually compared to the test case of clean speech only using a two sample statistical t-test with a 5% significance level and unequal variances. Again, the methods mitigate the train/test mismatch but are not statistically comparable to that of clean speech.

Table 4.6

ISR for comparison with clean speech

Test Speech	Approach	ISR	p-Value
Clean	MFCC	95.4	
G723.1	McCree + Affine, Borda Count	87.9	1.6e-07
G729	McCree + Affine, Borda Count	91.1	6.4e-05
G729	McCree with MFCC	91.1	1.06e-04
GSM-AMR	Affine Transform, Decision Level	89.8	1.52e-07

Note: Each ISR is in the form of an average percentage over 10 trials for a given condition. Two approaches from G729 that resulted in an identical ISR are included. McCree + Affine refers to using the combination of both enhancement methods.

Table 4.7

EER for comparison with clean speech

Test Speech	Approach	EER	p-Value
Clean	Product Fusion	2.77	
G723.1	McCree + Affine, Sum Fusion	4.10	2.3e-05
G729	McCree + Affine, Sum Fusion	3.13	0.02
GSM-AMR	Affine Transform, Product Fusion	3.24	9.9e-04

Note: Each EER is in the form of an average percentage over 10 trials for a given condition. McCree + Affine refers to using the combination of both enhancement methods.

Chapter 5

Conclusions

This chapter details a final discussion and the conclusions of this thesis. A review of the purpose and scope of the thesis is discussed. A complete list of the research accomplishments of this thesis is provided. Recommendations for the research and for potential future work is also discussed.

5.1 Thesis Review

The first chapter is an introduction to a speaker recognition system and the problem that speech coding distortion presents. The second chapter provides in depth background information for all aspects of the speaker recognition systems which include the system initialization, implementation, testing, and statistical analysis. All related derivations and equations related to this background information are provided in this chapter. The third chapter details the approach and methodology for training and testing the speaker recognition systems. The fourth chapter contains the complete results of the extensive testing performed using the aforementioned approach. A statistical analysis is also performed in order to prove that the results obtained are statistically significant.

5.2 Research Accomplishments

The purpose of this thesis was to research, develop, and implement a novel enhancement method to mitigate the negative performance effects of speech coding distortion on a speaker recognition system. The results showed that the use of the affine transform provided a statistically significant improvement of system performance when the enhancement method was applied to a speaker recognition system. The objectives as

described in the first chapter are restated and the research accomplishments of this thesis are examined below:

1. *To improve the performance of a speaker recognition system by reducing the effect of speech coder distortion.*

- A software (MATLAB) based speaker identification system (SI) and speaker verification system (SV) is designed and implemented. Four features are used for both the SI and SV systems which include the cepstrum (CEP), adaptive component weighting (ACW), postfilter cepstrum (PST), and mel-frequency cepstral coefficients (MFCC). The MFCC feature is generally the optimum feature. Each type of coder distortion G723.1 (6.3 kbps), G729 (8 kbps), and GSM AMR (12.2 kbps) affect the classification ability of the features.

2. *To implement a GMM-UBM based system.*

- A Gaussian mixture model universal background model based SI and SV system is implemented using various numbers of mixtures (16 to 2048 in powers of 2). The adaption of the weights, means, and covariances as well as just adapting the means only for each of the four features is also performed. A corresponding UBM for each feature is developed.

3. *To implement feature enhancement by applying the affine transform*

- The affine transform is the novel feature enhancement method proposed and implemented in this thesis.

4. *To implement signal enhancement by applying the McCree method.*

- The signal enhancement method (McCree method) performs better than feature enhancement (affine transform) on the lower bit rate G729 and G723.1 coders. Feature enhancement performs better on the higher bit rate GSM AMR coder.

5. *To combine feature and signal enhancement.*

- Both the feature (affine transform) and signal (McCree method) enhancement strategies are highly useful in improving the performance of SI and SV systems that are trained on clean speech and tested on the decoded speech. The combination approach is optimum for the lower bit rate G723.1 and G729 coders. Feature fusion (affine transform) is the optimum enhancement method for the higher bit rate GSM AMR (12.2 kbps) coder.

6. *To implement post-processing fusion techniques to further augment performance.*

- Feature based fusion methods for the SI system include decision level fusion and Borda count method. Both feature fusion methods do not improve performance for the lower bit rate G723.1 and G729 coders. Decision level fusion performs better for the higher bit rate GSM AMR coder while the Borda count method does not.
- Score fusion methods for the SV system include sum, product, and maximum fusion. The difference in performance of sum and product score fusion methods when compared to the MFCC feature is not statistically significant for all three coders. Sum and product fusion perform better than maximum fusion for G729 and GSM AMR but not better for the lowest bit rate G723.1 coder.

7. *To determine the optimal set of system parameters for the implementation of a speaker recognition system. These parameters include the number of Gaussian mixtures, the speech features used, the type of enhancement method and the fusion strategy.*

- The use of 256 mixtures and adapting means only was experimentally found to be the optimum parameter set. This narrowed approach allowed for a total of ten unique trials to be performed for each feature, each enhancement method, and each fusion method.

8. *To apply statistical techniques to compare the different approaches to determine statistical significance.*

- A two-way analysis of variance (ANOVA) provides the statistical proof necessary to decide which approaches perform better than others. A two-sample t-test allows statistical comparison of the final optimal approaches on speech with coding distortion to be compared with the optimal clean speech benchmark.

5.3 Research Recommendations and Future Work Considerations

The approaches in this thesis have been exhaustively tested in regards to mitigating speech coding distortion. Additional variables such as additive noise in combination with speech coding distortion could also be investigated. Additional classifiers for the purposes of classifier based fusion in a further attempt to mitigate speech coding distortion can be investigated. The addition or removal of certain features can be explored. The use of different speech coders especially those of higher bit rates

can also be considered to see if the enhancement methods that are proposed are still as effective.

References

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, 2008.

[2] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues", *IEEE Circuits and Systems Magazine*, pp. 23–61, June 2011.

[3] H. Beigi, *Fundamentals of Speaker Recognition*, Springer, 2011.

[4] "ITU-T: Recommendation G.723.1 - Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s", 1996.

[5] P. Kabal, "ITU-T G.723.1 Speech Coder: A Matlab Implementation", Telecommunications and Signal Processing Laboratory, McGill University, 2004.

[6] "ITU-T: Recommendation G.729 - coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear prediction (CS-ACELP)", 2007.

[7] "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech CODEC speech processing functions; AMR speech CODEC; General description", 2012.

[8] M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 3, pp. 260–267, May 1998.

[9] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust Speaker Recognition; A Feature-based Approach", *IEEE Signal Processing Magazine*, September 1996.

[10] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, “A Fast Algorithm for Finding the Adaptive Component Weighed Cepstrum for Speaker Recognition”, *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 1, pp. 84–86, January 1997.

[11] K. Raval, R. P. Ramachandran, S. S. Shetty and B. Y. Smolenski,“Feature and Signal Enhancement for Robust Speaker Identification of G.729 Decoded Speech”, *International Conference On Neural Information Processing, Doha, Qatar*, pp. 345–352, November 2012.

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

[13] A. McCree, “Reducing speech coding distortion for speaker identification”, *IEEE Int. Conf. on Spoken Language Processing*, 2006.

[14] U. Bhattacharjee and K. Sarmah, “GMM-UBM Based Speaker Verification in Multilingual Environments”, *International Journal of Computer Science Issues*, Vol. 9, No. 2, pp. 373-380, November 2012.

[15] F. Rastoceanu and M. Lazar, “Score fusion methods for text-independent speaker verification applications”, *6th Conference on Speech Technology and Human Computer Dialogue*, 2011.

[16] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Brooks/Cole Cengage Learning, 2012.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[18] I. T. Nabney, *NETLAB: Algorithms for Pattern Recognition*, Springer, 2002.

[19] A. Fazel and S. Chakrabartty, “An overview of statistical pattern recognition techniques for speaker verification” *IEEE Circuits and Systems Magazine*, pp. 62–81, June 2011.

[20] T. Hasan and J. H. L. Hansen, “A study on universal background model training in speaker verification”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 7, pp. 1890–1899, September 2011.

[21] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors”, *Speech Communication*, vol. 52, pp. 12–40, 2010.