Speech Analysis using Relative Spectral Filtering (RASTA) and Dynamic Time Warping (DTW) methods

by

Muhammad Amirul Azzim bin Zulkifly

18313

Dissertation submitted in partial fulfilment of
the requirements for the
Bachelor of Engineering (Hons)
(Electrical and Electronics)

JANUARY 2017

Universiti Teknologi PETRONAS
32610 Bandar Seri Iskandar
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Speech Analysis Using Relative Spectral Filtering (RASTA) And Dynamic Time Warping (DTW) Methods

by

Muhammad Amirul Azzim bin Zulkifly
18313

A project dissertation submitted to the
Electrical and Electronics Engineering Programme
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
BACHELOR OF ENGINEERING (Hons)
(ELECTRICAL AND ELECTRONICS)

Approved by,

____

(Dr. Norashikin binti Yahya)

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR, PERAK

January 2017

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

MUHAMMAD AMIRUL AZZIM BIN ZULKIFLY

ABSTRACT

This work consists of analysis of speech using RASTA and DTW methods. The analysis is based on the speech recognition. Speech recognition converts identified words or speech in spoken language into computer-readable format. The first speech recognition has been developed in the year of 1950s. The variation of speech spoken by individual becomes the main challenge for the speech recognition. Speech recognition has application in many areas such as customer call centers and as a medium in helping those with learning disabilities. This work presents an analysis of speech for Malay single words. There are three stages in speech recognition which are analysis, feature extraction and modeling. The Relative Spectral Filtering (RASTA) is used as the method for feature extraction. RASTA is a method that subsidized the undesirable and additive noise in speech recognition. Dynamic Time Warping (DTW) method is used as the modelling technique. DTW computes the optimal warping trajectory between two signals. The Singular Value Decomposition (SVD) technique is used to find the suitable eigenvector to use in comparing the signals. The signals compared is based on the average distance value generated using RASTA and DTW method. Experiments to determine the similarity of different signals are conducted and the results are expressed in terms of the distance value of comparing signals. The performance of speech processing techniques based on MFCC, RASTA and STFT will be compared. The trends of distance values by using different eigenvectors also will be shown.

ACKNOWLEDGEMENT

I would like to take this opportunity to give my sincere appreciation towards the persons and organizations that have been directly or indirectly provide contributions towards the completion of my Final Year Project to success.

I would like to express my greatest appreciation to my Final Year Project supervisor, Dr. Norashikin binti Yahya, for her patience, guidance and supervision throughout the completion of this project. This project would not be able to complete in time without her continuous encouragement and endless guidance.

My gratitude also goes to the lecturers, GAs, staffs and colleagues for their direct and indirectly involvement towards the completion of this project.

Last but not least, I would like to express my deepest appreciation to my parents and family members for their continuous support during completing this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND NOMENCLATURES

RASTA	Relative Spectral Filtering
DTW	Dynamic Time Warping
SVD	Singular Value Decomposition
STFT	Short Time Fourier Transform
WER	Word Error Rate
DWT	Discrete Wavelet Transform
LPC	Linear Predictive Coding
MFCC	Mel-Frequency Cepstral Coefficients
PCA	Principal Component Analysis
LDA	Linear Discriminate Analysis
ICA	Independent Component Analysis
MFFC	Mel-Frequency Cepstrum
HMM	Hidden Markov Model
PLP	Perceptual Linear Predictions

CHAPTER 1

INTRODUCTION

1.1 Background

Speech is a way of communication or transmit information with one another in which the receiver will receive the information and interpret it to give necessary responds. Speech recognition is a process of recognizing phrases or words in spoken language and transform them to a form that understandable by computers. Speech recognition has been developed throughout the decades. In the early 1950s, the innovation of “an isolated digit Recognition system for a single speaker” by Davis, Biddulph and Balashek at Bell Laboratories were the first attempt to create the speech recognition system [1]. A lot of researches have been done on the methods and techniques to enable computers to record and interpret human speech since the 1960s. By the 1980s, the first speech recognition system which could actually interpret speech was created [1, 2].

There are abundant of applications by speech-recognition technology such as a customer call centers’ routing system which activated by voice and voice dialing on mobile phones as well as a medium to help individuals with physical disabilities. Although there is no valid proof on the impacts to the students with physical disabilities, speech recognition software can be the best alternative to help those students with difficulties in writing fluency [3]. Crucial characteristics of robust speech recognition systems include noise filter and able to handle acoustic conditions such as the speaker’s speech rate and accent. The speech recognition system involves three main stages which are analysis, feature extraction and speech modeling. In this work, an isolated work recognition algorithm will be developed using dynamic time warping. The evaluation of speech recognition system performance will be based on its accuracy, rated using word error rate (WER) and processing time.

1.2 Problem Statement

The major challenge in analyzing speeches is the speaking style by the speakers. All humans have differences in speaking. Each person has their own accents and unique way to pronounce and emphasize. Humans also portray their emotions through speech. We speak differently depends on our current emotions such as sad, stressed, happy, disappointed etc. Besides the speaking style, the noise condition also plays a vital role in making speech analysis difficult. The high level of noise will increase the disturbance in the analysis process of the speech.

In this work, Malay language has been identified as the language to be used. Speech recognition algorithm using RASTA and DTW will be analyzed in recognizing some simple Malay words. Training data will be collected from several speakers with various types of speech such as style, emotion and accent. The data will be taken at different time as a means to introduce variability in the speech.

1.3 Objective and Scope of Study

The main objective of this study is:

- To analyze speech recognition algorithm using RASTA and Dynamic Time Warping method.

The scope of this study includes clear understanding on the overall system of speech recognition. Thorough analysis needs to be done on the techniques of speech recognition especially on the dynamic time warping method. The analysis on types of signals is an essential to be focused on as this study will involve a lot of signal observations and analysis. This study also involves MATLAB software as the platform in analyzing and simulating the speech recognition algorithm.

CHAPTER 2

LITERATURE REVIEW

2.1 Speech Recognition Architecture

Speech recognition is made up of three primary stages which are speech detection as the first stage, followed by feature extraction and lastly the pattern comparison or matching. A general speech recognition system block diagram is shown in figure 1.

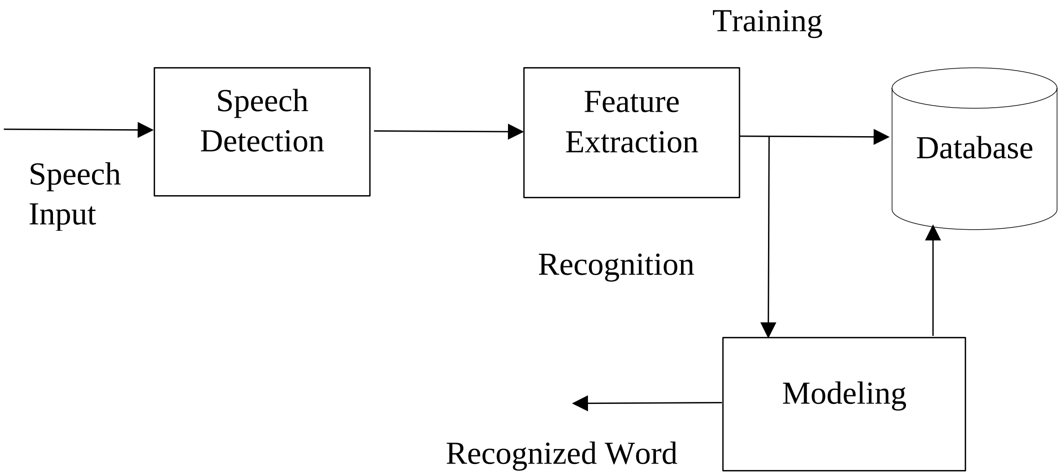


Figure 1: General 6Speech Recognition Block Diagram [4]

2.1.1 Speech Preprocessing

The speech recognition starts with the preprocessing or speech detection stage. At first, speaker's voice will be detected and transform into speech signal as the input. The input then will go through preprocessing stage. In this stage, all the disturbances for the speech such as the speaking style differences of the individual and noise will be removed. This removal action is done by preprocessing that involves anti-aliasing bandpass filter and voice sampling [4-6].

2.1.2 Feature Extraction

The second stage called feature extraction stage is where some of training elements are extracted from the speech signal. The training elements that will be generated contain essential information of the speech [4-7]. There are numerous techniques on feature extraction method. The most common techniques for feature extraction are the discrete wavelet transforms (DWTs), Linear Predictive Coding (LPC) and the Mel-Frequency Cepstral Coefficients (MFCC) [4, 6]. Table 1 shows the list of methods for feature extraction and their properties.

Table 1: List of Method for Feature Extraction and their properties [1]

No.	Method	Property and Implementation
1	Principal Component Analysis (PCA)	Indirect feature extraction method, good for Gaussian data
2	Linear Discriminate Analysis (LDA)	Indirect feature extraction method, better in classification than PCA
3	Independent Component Analysis (ICA)	Indirect feature extraction method, utilized for de-mixing non-Gaussian distributed features
4	Linear Predictive Coding (LPC)	Fixed feature extraction method, used in feature extraction for lower order
5	Cepstral Analysis	Fixed feature extraction method, used in representing spectral envelope
6	Mel-frequency scale analysis	Fixed feature extraction method, Spectral analysis
7	Filter bank analysis	Filters required frequencies
8	Mel-frequency Cepstrum (MFFCs)	Power spectrum is found by using Fourier Analysis, used to find the features required

9	Kernel based feature extraction method	Indirect transformations, used in repetitious features and for classification error upgrade
10	Wavelet	Better in time resolution compared to Fourier Transform
11	Mel-Frequency Cepstral Coefficients (MFCC)	Mimic human auditory system, helps in reducing frequency information of input speech signal into coefficients
12	Spectral subtraction	Robust feature extraction method, used basis on Spectrogram
13	Cepstral mean subtraction	Robust feature extraction, similar to MFCC but using the mean statically parameter
14	RASTA filtering	Widely used method for signals that have environmental noise or speech with noisy disturbance.
15	Integrated Phenome subspace method (Compound Method)	Transformation based on the sum of PCA, LDA and ICA

Based on table 1, each of the methods for feature extraction has its own advantages and disadvantages as well as their specific role in extracting features. We can see that there are some methods that can be classified into types of feature extraction method such as nonlinear, static, and robust feature extraction methods. Among all methods, MFCC is the best method for extraction method as MFCC mimics the human auditory system which means it is accurate and reliable in finding the required features.

2.1.3 Modeling

For the last stage, it is the modeling stage. In this stage, the aim is to produce speaker models by using speaker actual characteristic vector. For modeling technique, the computer supposed to abandon the characteristic of the speech signal such as the noise, accent, style and only separate the desired message [1]. There are many methods fall under modeling technique.

Table 2: Modeling Techniques [1]

No.	Techniques	Properties
1	Acoustic-phonetic approach	Based on finding speech sounds and specify relevant labels to these sounds, not been widely used
2	Pattern Recognition approach • Hidden Markov Model (HMM)	Use a formulated mathematical approach, forms persistent speech pattern for reliable comparison from the speech database
3	Template based approaches	Matching unknown speech signal with collection of prototypical speech patterns, able to avoid errors made by segmentation of smaller acoustic units
4	Dynamic Time Warping	An algorithm in measuring the correlation between two sequences in terms of time or speed, reliable in recognize isolated word
5	Knowledge based approaches	Uses the knowledge of experts about variations in speech
6	Statistical based approaches	Speech variations are modeled analytically
7	Stochastic Approach	Using the probabilistic models in order to handle insufficient information
8	Artificial intelligence approach	Hybrid of the acoustic-phonetic and pattern recognition approach

Table 2 shows the various modeling techniques with their pros and cons. Based on the table, there are different approach in each of the methods in order to find the real match in recognizing speech. Pattern recognition approach can be said as the best techniques in modeling stage of speech recognition because of the Hidden Markov Model (HMM) techniques which uses formulated mathematical approach. By using this mathematical approach, it will result in accurate matching of the unknown speech with the database.

2.2 Speech Recognition Techniques

Speech recognition has been developed with many kinds of methods throughout the centuries. There are many kinds of technologies and approaches have been done to improve and innovate the speech recognition. Some methods that are widely used in the speech recognition are the Hidden Markov Model (HMM), Mel-frequency cepstral coefficients (MFCC), Linear Predictive Coding (LPC) and Dynamic Time Warping.

2.2.1 Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is used as one of the techniques in modeling of speech recognition system. In today’s era, the dependency on modeling of acoustic unit, which is the speech units in the speech recognition system is high for most speech recognition systems. We used these acoustic units to derive a model that will be used to recognize non-stop speech signal. In this matter, HMM is widely used to be the suitable model. HMM is very often used because of it allows around 80% of recognition for a given speech signal, however the rate is still not desirable [8]. HMM is widely chosen for speech recognition because it gives simplicity and flexibility in mechanism for modeling sequences [9].

2.2.2 Mel-Frequency Cepstral Coefficients (MFCC)

There many methods for feature extraction stage in speech recognition. Mel-frequency cepstral coefficients (MFCC) is the preferable method for feature extraction of speech recognition [4, 6, 10]. This method is depending on the perception of human's hearing system [6, 10, 11]. The advantages of MFCC is that it is less sensitive to noise due to MFCC produce training vectors by changing the speech signal into frequency domain [4]. Chronology on how MFCC system works is shown in figure 2.

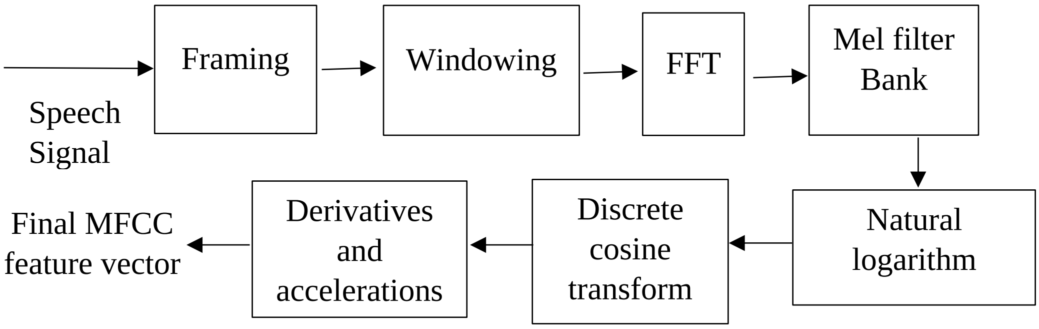

Speech Signal
Framing
Windowing
FFT
Mel filter Bank
Natural logarithm
Discrete cosine transform
Derivatives and accelerations
Final MFCC feature vector

Figure 2: MFCC Block Diagram [11]

2.2.3 Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) is a method of compression and synthesis of speech recognition. LPC highly aims on producing the vectors that can be observed and extracting the vocal tract parameters [12, 13]. LPC is better worked in clean environment despite in noisy situation. LPC is the key role in the usage of auditory and processing speech signal. LPC consists of pre-emphasis, framing, windowing and computing which will be shown in figure 3.

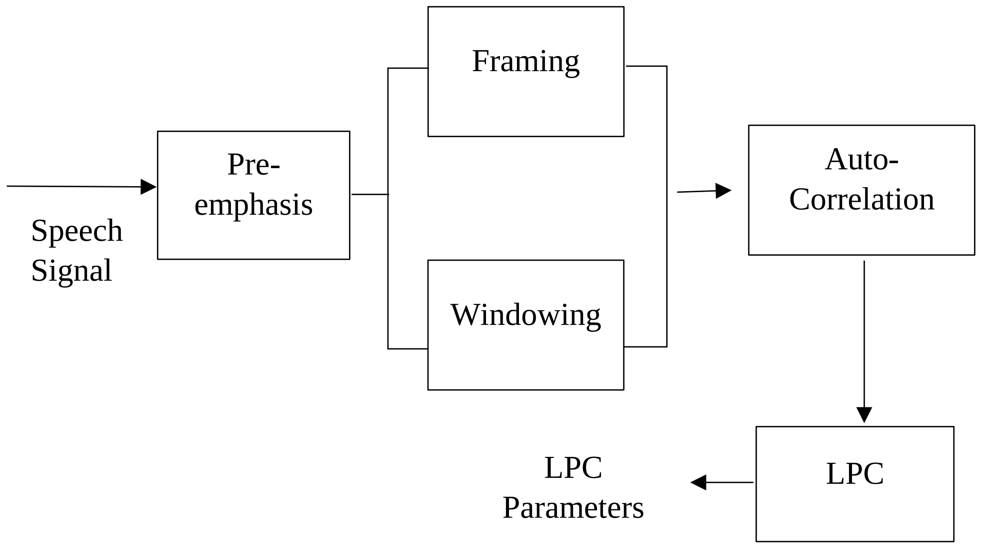


Figure 3: Linear Predictive Coding Block Diagram [14]

2.2.4 Dynamic Time Warping

Dynamic Time Warping or DTW is used in the modeling stage of speech recognition. DTW is an algorithm to do comparison between two sequences that based on time and speed [1, 4, 15]. DTW aims to put two sequences of feature vectors in parallel by using the technique of warping the time axis iteratively until an ideal match between the two sequences is reached. DTW is good to be used in a system which recognized isolated word and also performs well in recognizing connected word.

2.3 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a principle for eigenvalues and eigenvectors in representing a square matrix. SVD is best known as the significant tools of numerical linear algebra. Information about the rank of the matrix, noise level and the energy is equipped by the singular value of a given data matrix. This singular value is also produce the only computationally dependable method for computing ranks of operators and chosen fundamental objects of linear algebra [16]. The SVD technique plays a major role in separating the actual signals from recorded data matrix because this technique is very robust and numerically dependable [16]. Based on many signal processing applications and control systems, the signal value decomposition of matrix formed from observed data can be used to enhance techniques of signal parameter estimation and system verification [17].

CHAPTER 3

METHODOLOGY

3.1 Research Methodology

Speech recognition consist of two main stages which are feature extraction and modeling stage. In this work, the technique for feature extraction is the Relative Spectral Filtering (RASTA) method and for modeling the technique that will be used is the Dynamic Time Warping method. The block diagram for the speech recognition system is shown in figure 4.

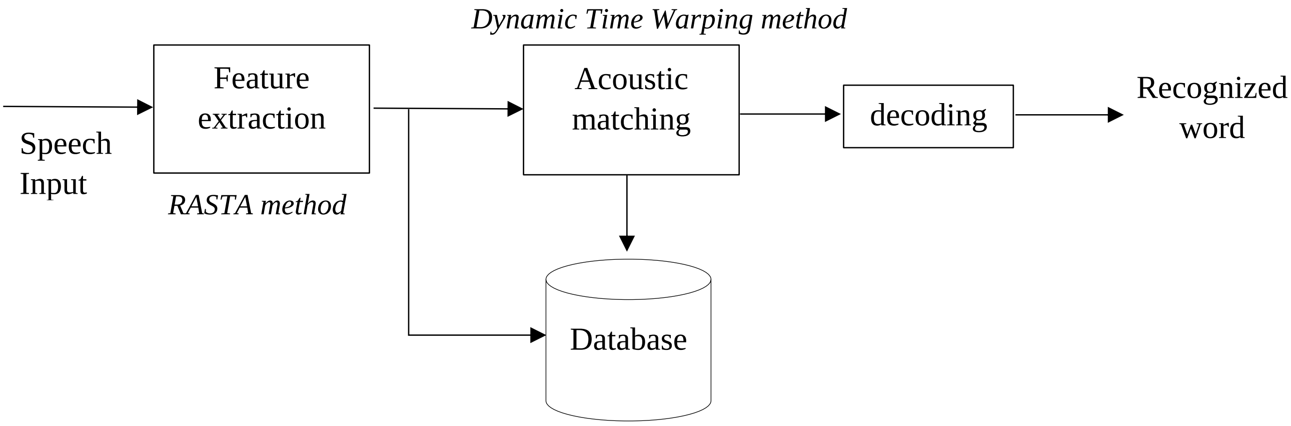


Figure 4: Speech Recognition block diagram

First of all, Malay words will be collected from variety of speakers and stored in the database. Speaker’s voice will be the speech input in form of signal. The input speech then will be cleaned up and filtered in the pre-processing block and the feature extraction which using the RASTA method. After that, the clean input speech will be compared to find match characteristics such as speech style, accent, and emotion from the database. Lastly, the speech characteristic will be removed and only intended message is taken in the modeling stage.

3.2 Relative Spectral Filtering (RASTA) method

Relative Spectral Filtering or RASTA method is used in feature extraction stage for the speech recognition. This method is widely performed for signals with environmental noise or speech signal that have noisy disturbance [12]. RASTA method is very efficient in capturing signals with low modulation. The main function of RASTA method is that it diminishes the undesirable and additive noise in speech recognition. Moreover, RASTA is not just lessening the noise impact on the speech signal, but it also improves the quality of the speech signal. The block diagram of RASTA is shown in figure 5.

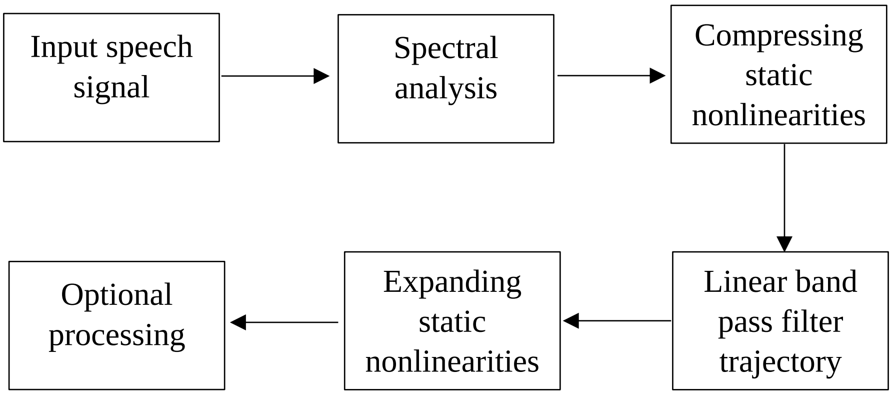

Input speech signal
Spectral analysis
Compressing static nonlinearities
Linear band pass filter trajectory
Expanding static nonlinearities
Optional processing

Figure 5: RASTA method block diagram [12]

RASTA algorithm processing steps [18]:

i) Measure the short time power spectrum of windowed signal.

ii) Convert spectral amplitude through the compressing static nonlinearities block.

iii) Filter the time path of each converted spectral factor.

iv) Convert the filtered speech representation through the widening static nonlinearities conversion block.

3.3 Dynamic Time Warping (DTW) method

In the modeling stage of the speech recognition, the Dynamic Time Warping or DTW method is used. DTW is an algorithm in measuring the correlation between two sequences in terms of time or speed. It is very reliable in recognize isolated word, thus it is the perfect method to use in this project. DTW implements a piece wise consecutive mapping of time axis to align the two signals. The calculated local distance which is (L), of both signals is placed in a grid. The best alignment of two sequences becomes the trajectory through the grid, which reduce the total distance between the two signals [7].

Algorithm for calculation of DTW scores [7]:

i) The reference(x) and testing (y) feature vectors separated from speech signal is stored in separate matrices.

$$x = [x_1, x_2, \dots, x_n] \; and \; y = [y_1, y_2, \dots, y_n] \qquad (1)$$

ii) The local distance (L) is calculated using the Euclidean distance formula.

$$L_{xy} = |x - y| = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2]^{1/2} \qquad (2)$$

iii) The global distance (G) is calculated using recursive formula.

$$G_{xy} = L_{xy} + \min\left(G_{x-1\,y-1}, G_{x-1\,y}, G_{x\,y-1}\right) \qquad (3)$$

Based on figure 6, the speech signals of single word database are read and preprocessing and feature extraction is executed on it. Then, the parameters that kept in vectors form are classified as training data. The score of DTW between testing speech signal and the signals in the database is calculated. The two signals is matched if the global distance between them is below the permissible limit or the shortest distances among all the speech sequences.

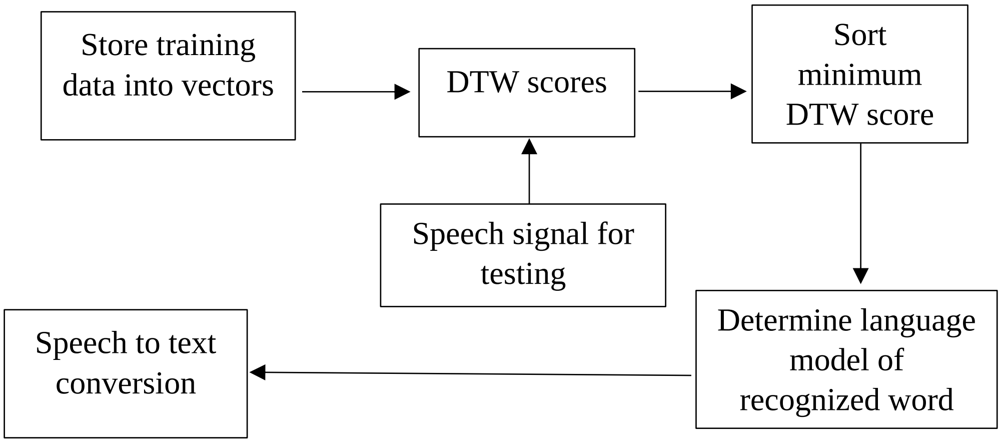

Store training data into vectors
DTW scores
Sort minimum DTW score
Speech signal for testing
Determine language model of recognized word
Speech to text conversion

Figure 6: DTW method block diagram [7]

3.4 The Singular Value Decomposition (SVD)

The following theorem shows the SVD for real matrices [19], [17]:

Theorem 1. For any real $m \times \ n\, matrix\ A,$ there exists a real factorization

$$A = U_{m\times m}\ \cdot\ S_{m\times n}\ \cdot\ V_{n\times n}^{T} \tag{4}$$

Where

$$U^T U = I_{m\times m} \tag{5}$$

$$V^T V = I_{n\times n} \tag{6}$$

In which the matrices U and V are orthogonal, and matrix S is real pseudo-diagonal with nonnegative diagonal elements.

Columns of U are the left singular vectors, S has the same dimension as A and has singular values. V^T has the right singular vectors rows. The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

Calculating SVD involved the findings of the eigenvalues and eigenvectors of AA^T and A^TA. The eigenvectors of A^TA make up the columns of V while the eigenvectors of AA^T make up the columns of U.

The singular values in S are square roots of eigenvalues from AA^T and A^TA. The singular values are the diagonal entries of the S matrix and are arranged in non-ascending order. The singular values are constantly real numbers. U and V are also if the matrix A is a real matrix.

Lemma 1. The number of non-zero singular values, equals the algebraic rank of the matrix A.

Lemma 2. $S = diag(\sigma_{1,}\sigma_{2,} \dots, \sigma_n)$ ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ (if σ is a singular value of A, its square is an eigenvalue of A^TA)

If $U = (u_1\, u_2\, \cdots\, u_n)$ and $V = (v_1\, v_2\, \cdots\, v_n)$, then

$$A = \Sigma_{i=1}^{r}(u_i \cdot \sigma_i \cdot v_i^T) \tag{7}$$

where $(u_i\, , \sigma_i\, , v_i\ \ )$ is the i-th singular triplet of matrix A.

Lemma 3. Frobenius norm of $m\ \times n$ matrix A of rank r

$$||A||_F^2 = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n}a_{ij}^2 = \Sigma_{k=1}^{r}\sigma_k^2 \tag{8}$$

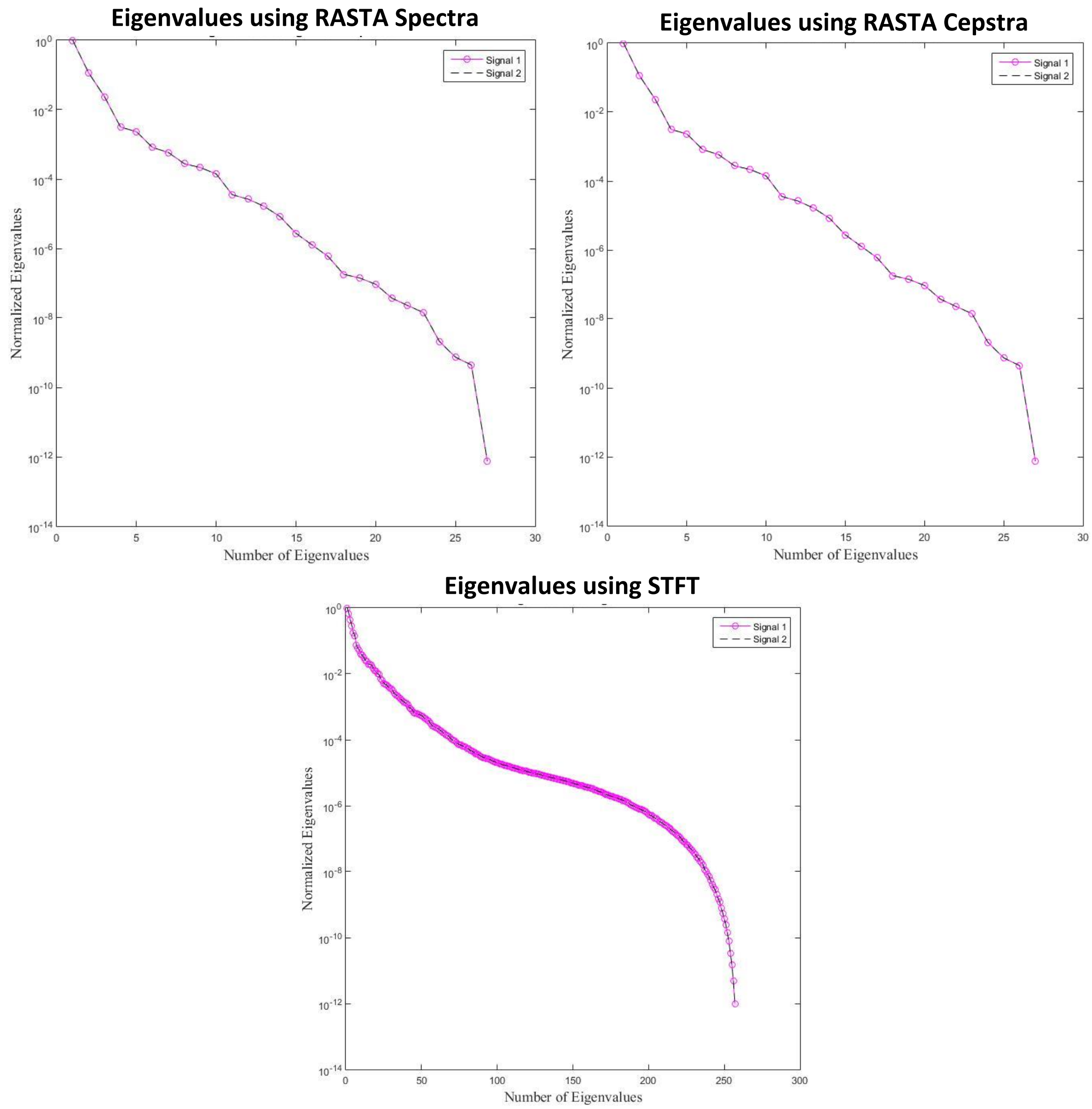
Where σ_k are the singular values of A.

The total energy in a vector sequence $\{a_k\}$ associated with matrix A as defined in definition 1, is equal to the energy in the singular spectrum.

The smallest non-zero singular value corresponds to the distance in Frobenius norm, of the matrix to the closest matrix of lower rank. This property makes SVD attractive for approximation and data reduction purposes.

3.4.1 Eigenvalues comparison between RASTA and STFT

Following figures will show the eigenvalues comparison between RASTA and STFT method. These figures aim to show the role of eigenvalues in differentiating between two different signals. Figure 7 shows the comparison between the same signal which is Malay word of ‘satu’ while figure 8 shows the comparison between two different signals which are Malay words of ‘satu1’ and ‘satu2’ produced by the same speaker with two different durations. ‘satu1’ has duration of 1 second while ‘satu2’ has duration of 2 seconds.

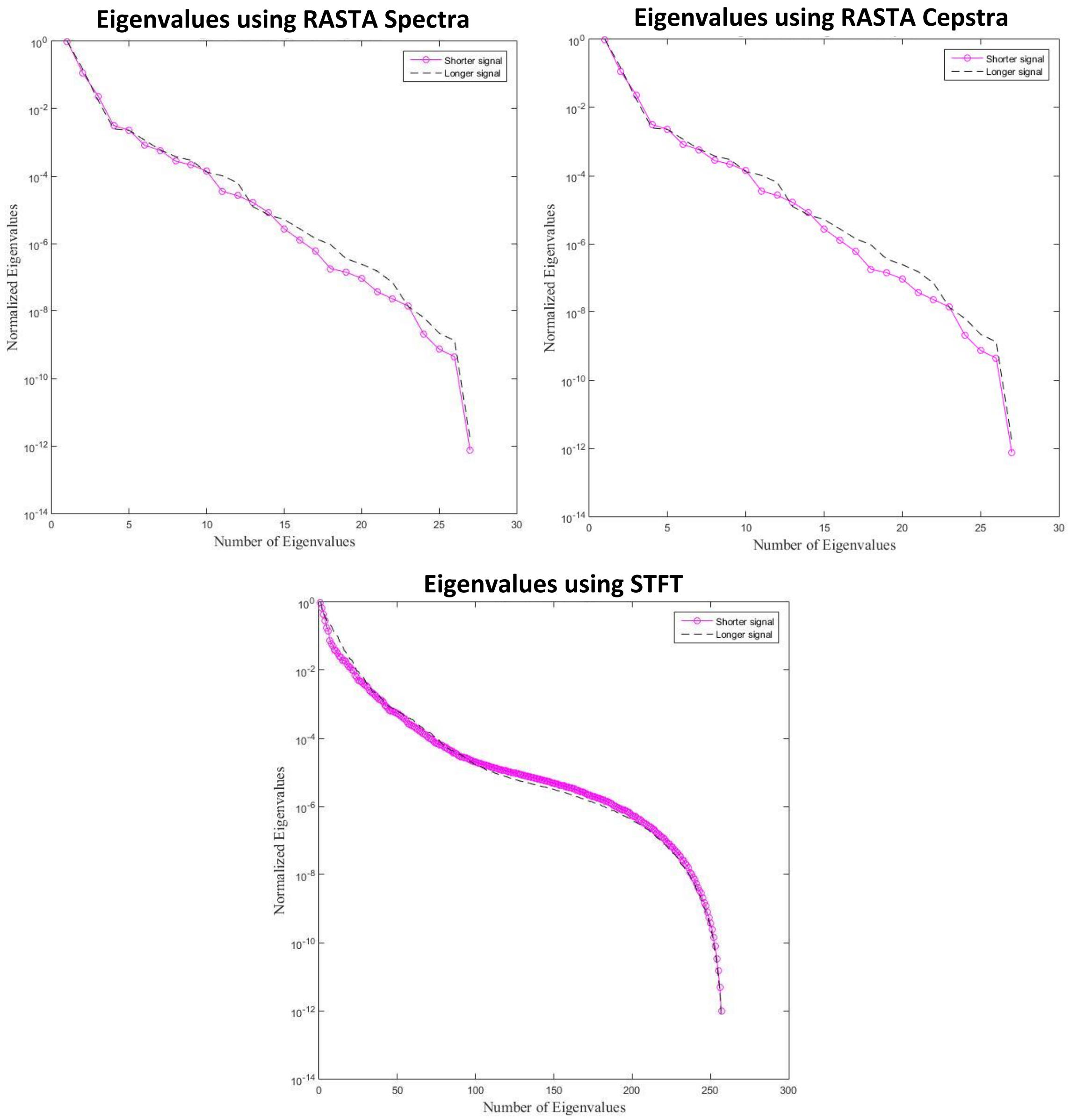


Eigenvalues using RASTA Spectra

Eigenvalues using RASTA Cepstra

Eigenvalues using STFT

Figure 7: Eigenvalues comparison by using RASTA and STFT between the same signal of same speaker



Eigenvalues using RASTA Spectra

Eigenvalues using RASTA Cepstra

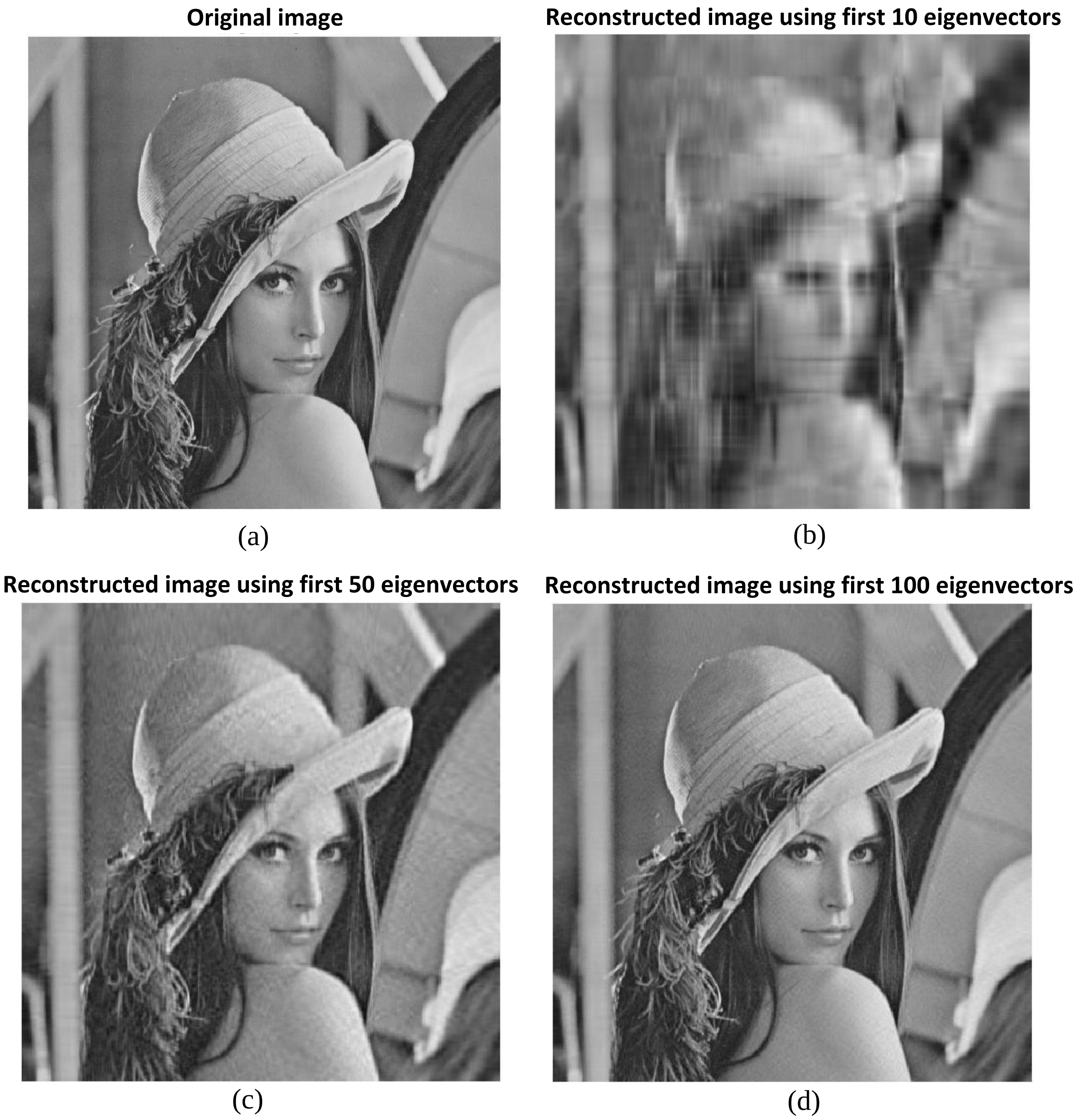
Eigenvalues using STFT

Figure 8: Eigenvalues comparison by using RASTA and STFT between signals ‘satu1’ and ‘satu2’ of same speaker

Based on figure 7, we can see that the circle shapes are constantly following the dashed line for both RASTA and STFT. This indicates that the two signals are perfectly match. On the other hand, figure 8 shows different results where the circle shape is no longer following the dashed line for both RASTA and STFT. This clearly indicates that the two signals are significantly different. Based on the two methods, we can see that RASTA shows clear difference between shorter signal and longer signal while STFT shows barely seen difference between shorter signal and longer signal.

3.4.2 Image compression using SVD

We can see properties of SVD in approximation and data reduction purposes by using image compression. Figures below will show the difference of Lena image between the original image and reconstructed image using some number of eigenvectors.



Original image

(a)

Reconstructed image using first 10 eigenvectors

(b)

Reconstructed image using first 50 eigenvectors

(c)

Reconstructed image using first 100 eigenvectors

(d)

Figure 9: Lena image differences between original and reconstructed image using first 10, 50 and 100 eigenvectors

Figure 9 shows the difference of Lena image between original and reconstructed image using first 10, 50 and 100 eigenvectors. Based on the figure, we can see that in (b), which is reconstructed image using first 10 eigenvectors produced a very significant difference compared to the original image. We can see that (b) is very noisy and blur. The details shown in (a) is cannot be seen completely. As we move on to (c), which is reconstructed image using first 50 eigenvectors, the image is begin to be less blur and shows sharper details of the Lena image. However, (c) is still has low similarity compared to the original image, (a). Lastly, in (d) which is the reconstructed image using first 100 eigenvectors, the image produced is clear and closely similar to the original image shown in (a). Based on this observation, as the first number of eigenvectors increased, the reconstructed image become closely similar to the original image.

3.5 Project Workflow

This speech recognition is done by stages. Figure 7 below shows the workflow in completing this project.

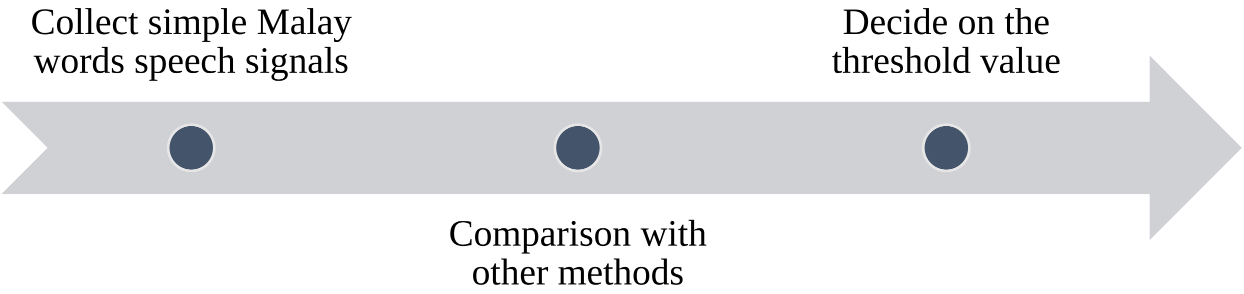

Collect simple Malay words speech signals
Decide on the threshold value
Comparison with other methods

Figure 10: Project Workflow

3.5.1 Collect simple Malay words speech signals

Samples of simple Malay words speech signals are collected. These signals are collected from a same speaker saying Malay words of ‘Satu’, ‘Dua’, ‘Tiga’, ‘Empat’, ‘Lima’, ‘Enam’, ‘Tujuh’, ‘Lapan’, ‘Sembilan’ and ‘Sepuluh’. Each word signal contains 20 different samples which comes from various speaking style and accents. This results to 200 samples of signals collected.

3.5.2 Comparison with other methods

Relative Spectral Filtering (RASTA) method is compared with another methods such as Mel-Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transform (STFT) in terms of RASTA’s ability to extract signal features and to produce good distance similarity values. From this comparison, we can evaluate the performance of RASTA method in the speech recognition.

3.5.3 Decide on the threshold value

The threshold value of the speech recognition system need to be determined as the reference for the distance similarity value produced by comparing the signals. This threshold value can be decided based on the distance similarity values pattern.

CHAPTER 4

RESULTS AND DISCUSSION

Experiments conducted to compare the distance of similarity between two different voice signals using several techniques, which are RASTA, Short Term Fourier Transform (STFT) and MFCC. Recorded voice of speaker saying single Malay words of 'SATU' with different way of pronunciations has been used as the voice signals to be compared for the experiment. The fist signal which is 'satu1' has the duration of 2 seconds while the second signal which is 'satu2' has longer duration which is 4 seconds.

4.1 Spectrogram of signals

Figures 11 and 12 show the pattern of both signals 'satu1' and 'satu2' respectively in terms of spectrogram. Based on the figures, we can see that the 2 signals look similar except there is temporal difference.

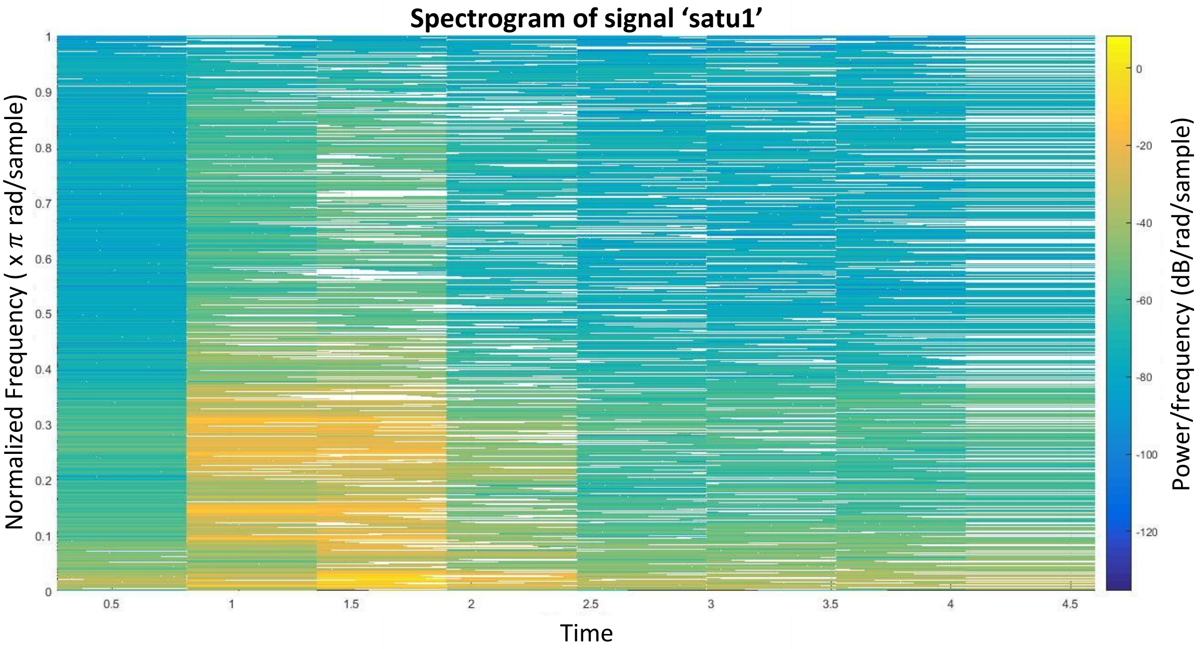


Figure 11: Spectrogram of 2 seconds duration voice signal 'satu1'

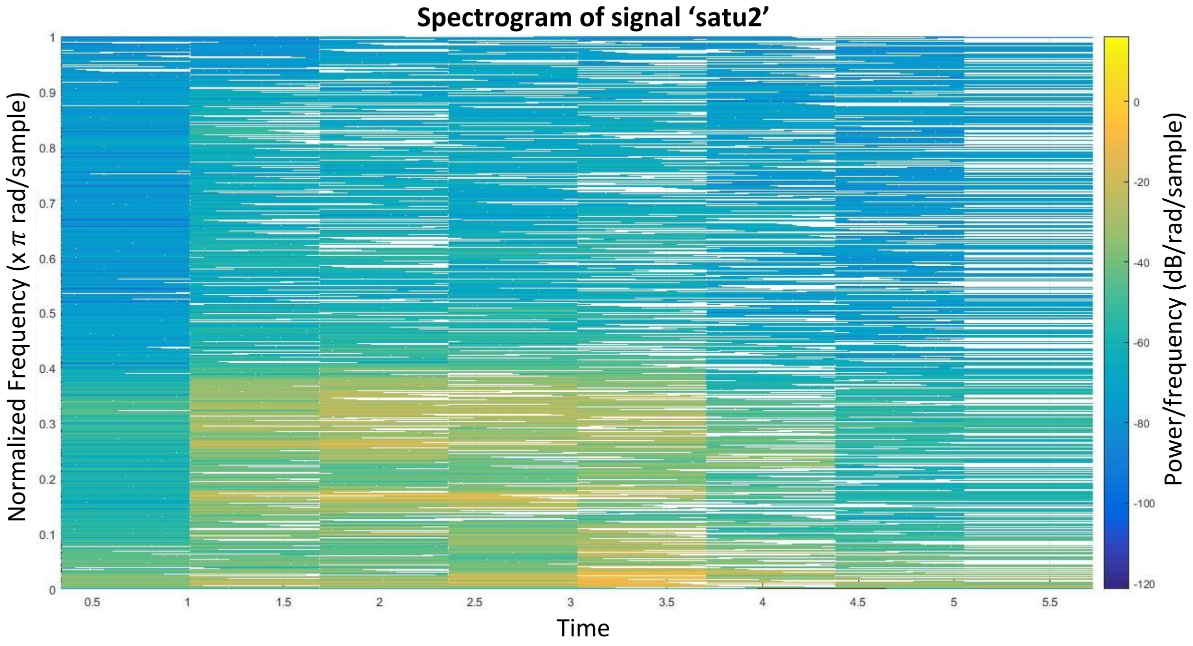


Figure 12: Spectrogram of 4 seconds duration voice signal 'satu2'

4.2 Extracted Signal Features using DTW distance, STFT, RASTA and MFCC

4.2.1 Word 'satu' by the same speaker

Figure 13 shows the spectral features of signals 'satu1' and 'satu2' by using RASTA-PLP while figure 14 shows the similarity matrix between signals 'satu1' and 'satu2' by using RATSA Spectra and Cepstra. Based on figure 13, we can see similar features produced by both signals in the beginning of each signal. Similar to figure 14 where it shows red straight line in the beginning of the similarity matrix. This red straight line indicates that both signals are perfectly match. Based on these figures, we can see that similar features by both signals can extracted and can be match even though both signals are different in terms of length. It shows that RASTA can extract the similar important information from both signals.

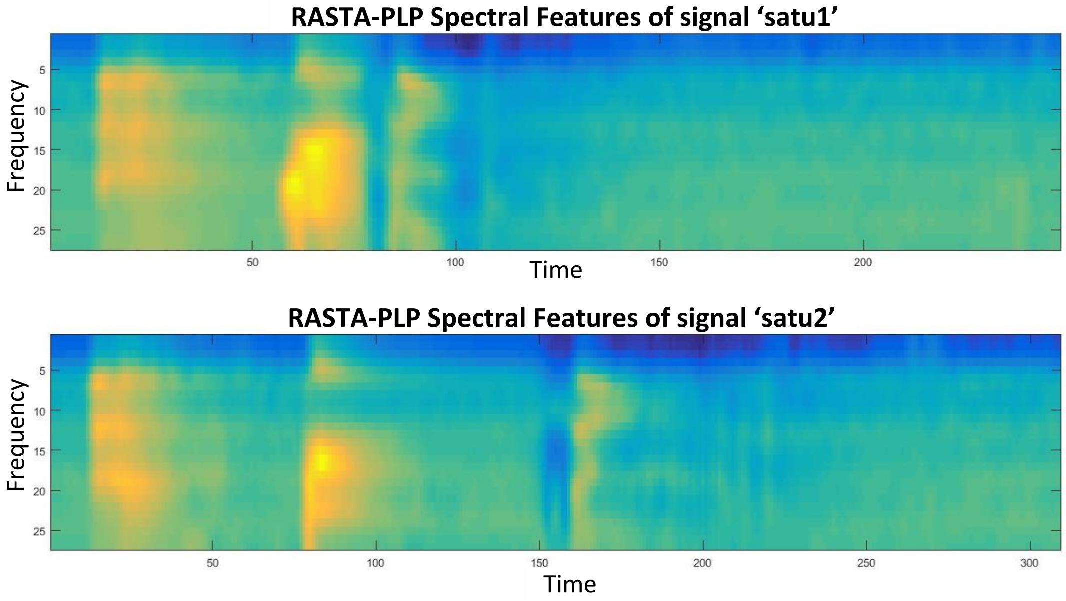


Figure 13: RASTA-PLP spectral features of signals ‘satu1’ and ‘satu2’ of same speaker

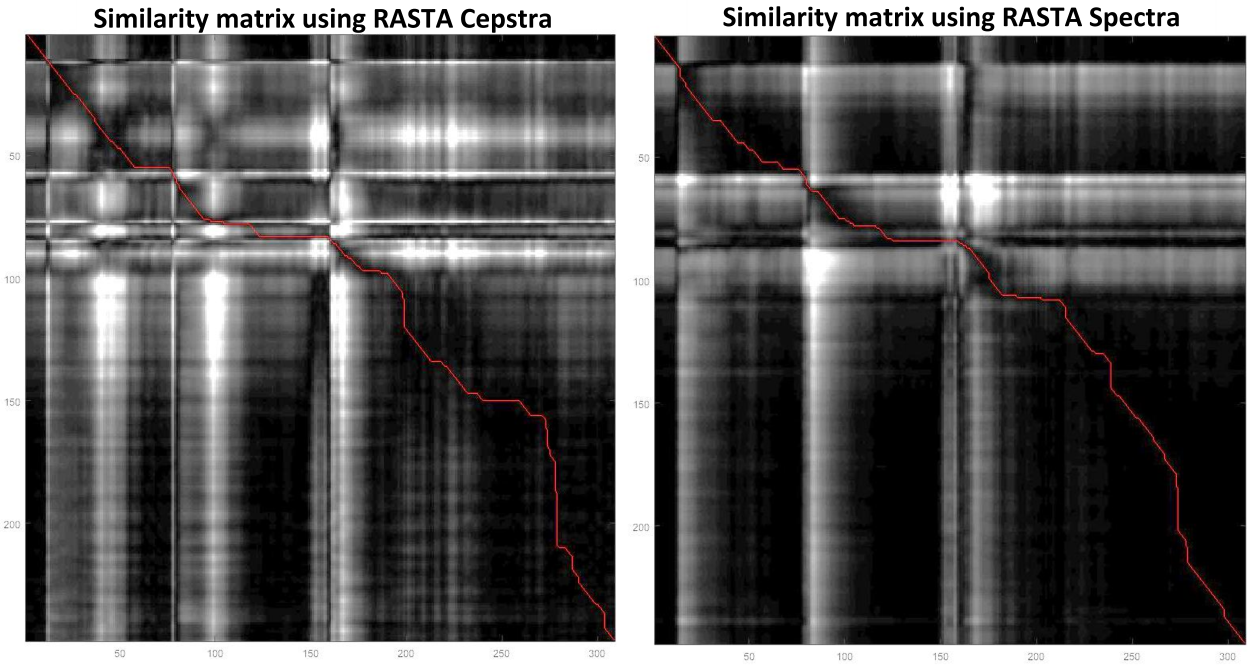


Figure 14: Similarity matrix between signals ‘satu1’ and ‘satu2’ of same speaker by using RATSA technique

Figure 15 shows the spectral features of signals ‘satu1’ and ‘satu2’ by using MFCC while figure 16 shows the similarity matrix between signals ‘satu1’ and ‘satu2’ by using MFCC. Similar with RASTA, MFCC also shows similar features produced by both signals. Based on figure 16, it also shows red straight line in the similarity matrix. This indicates that both signals in that time are perfectly match.

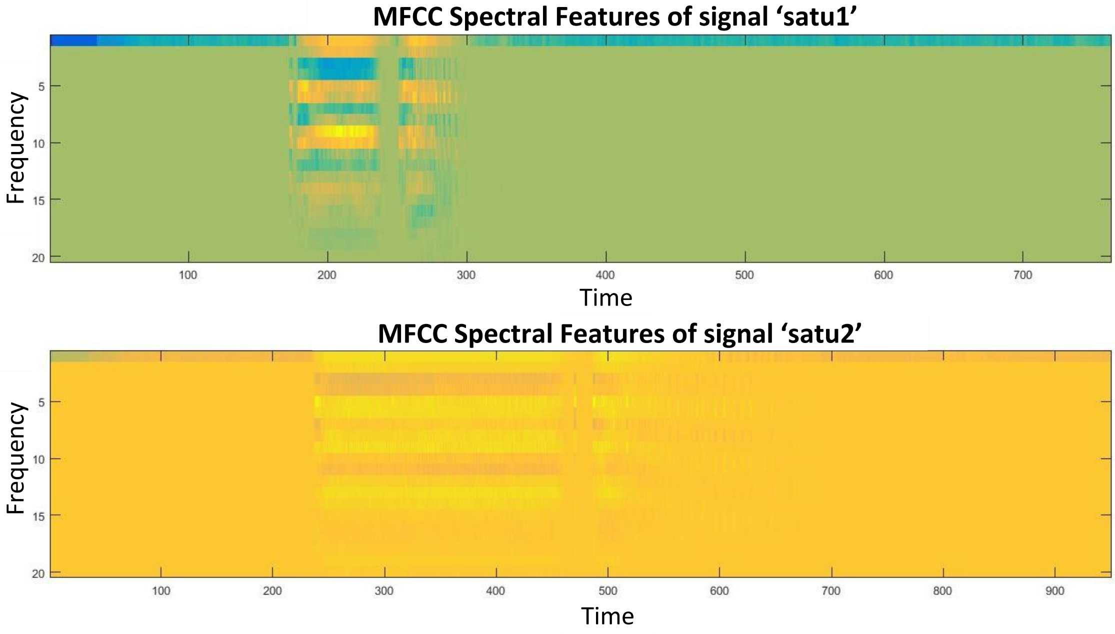
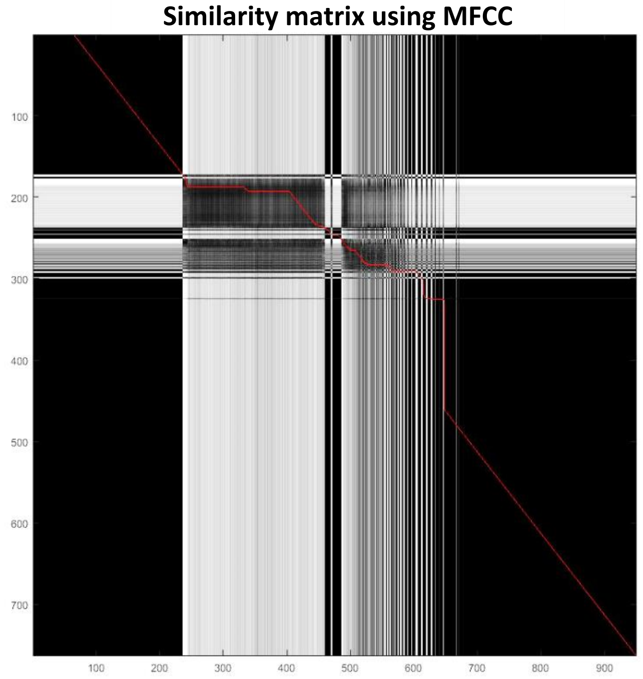


Figure 15: MFCC spectral features of signals ‘satu1’ and ‘satu2’ of same speaker



Similarity matrix using MFCC

Figure 16: Similarity matrix between signals ‘satu1’ and ‘satu2’ of same speaker by using MFCC technique

Figure 17 shows the similarity matrix of same signal which is signal ‘satu1’ by using different techniques which are STFT, RASTA Spectra, RASTA Cepstra and MFCC. This similarity matrix shows the correlation or in other words the similarity of the two signals in terms of distance. The red line illustrates an optimal cost path from the starting to the end of both the signals. We can see that all the similarity matrix show red diagonal straight line which indicates the both signals compared are perfectly match and the distance between the two signals is zero.

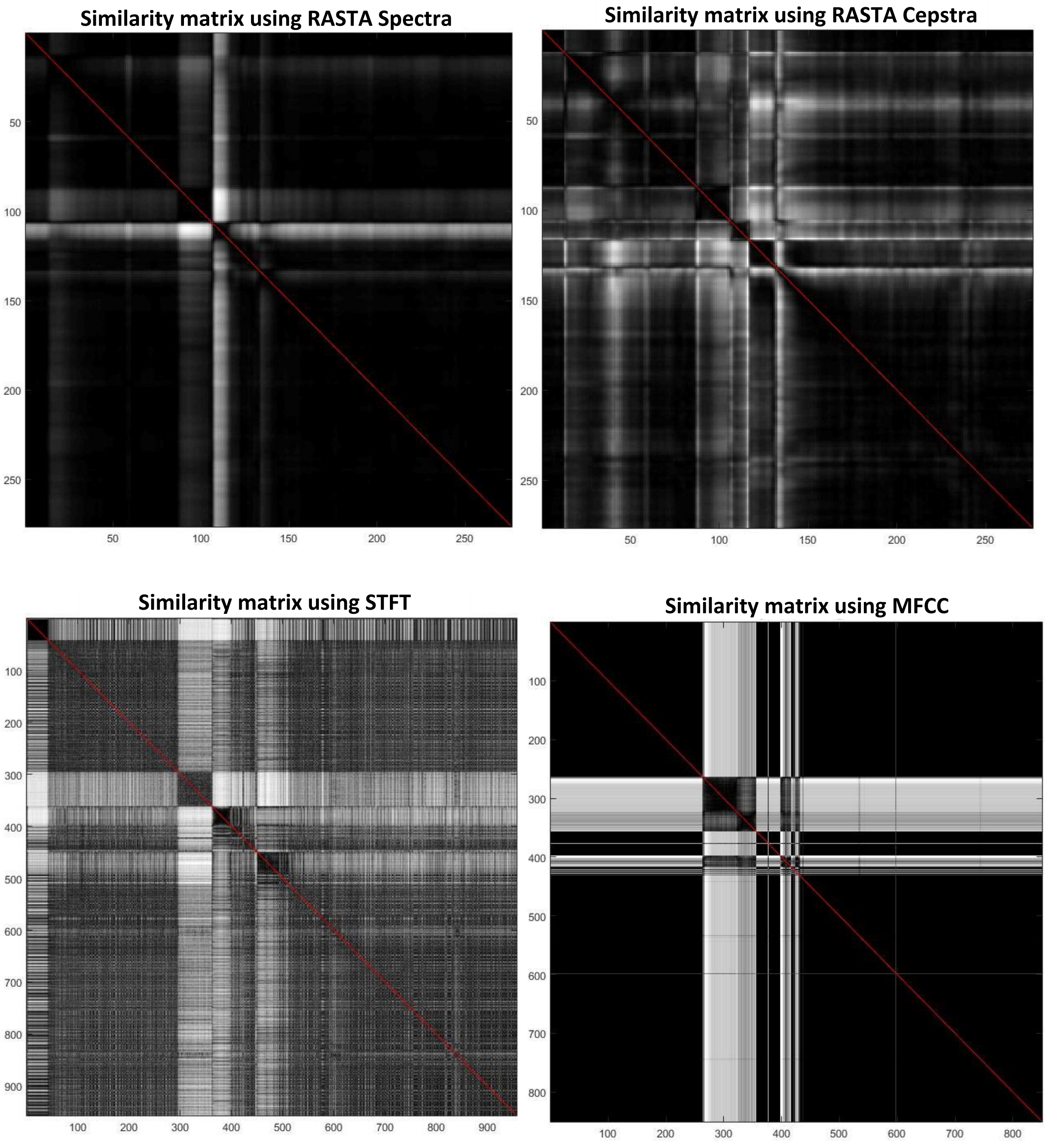
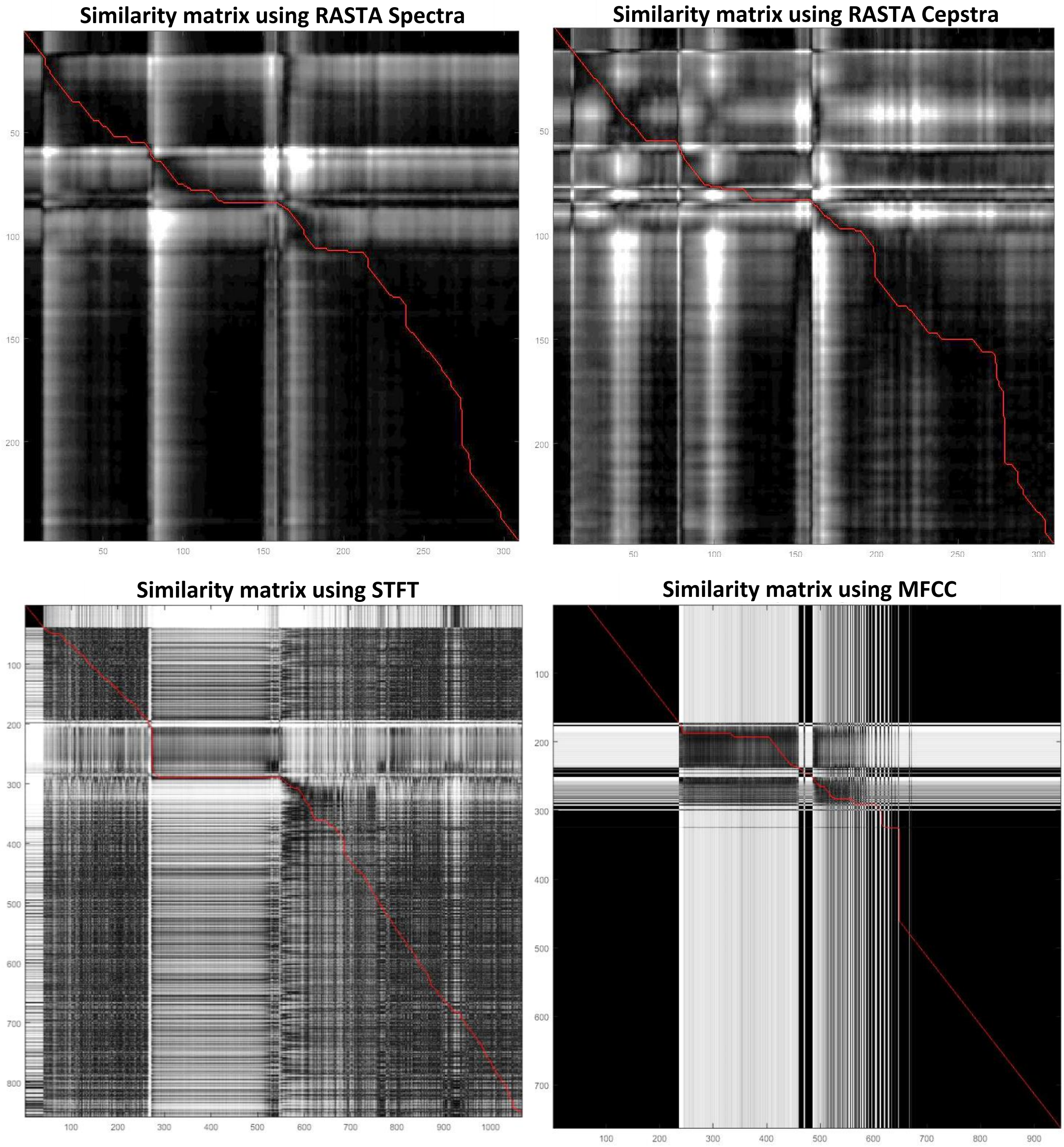


Figure 17: Similarity matrix by using different techniques between same signal ‘satu1’ of same speaker



Similarity matrix using RASTA Spectra

Similarity matrix using RASTA Cepstra

Similarity matrix using STFT

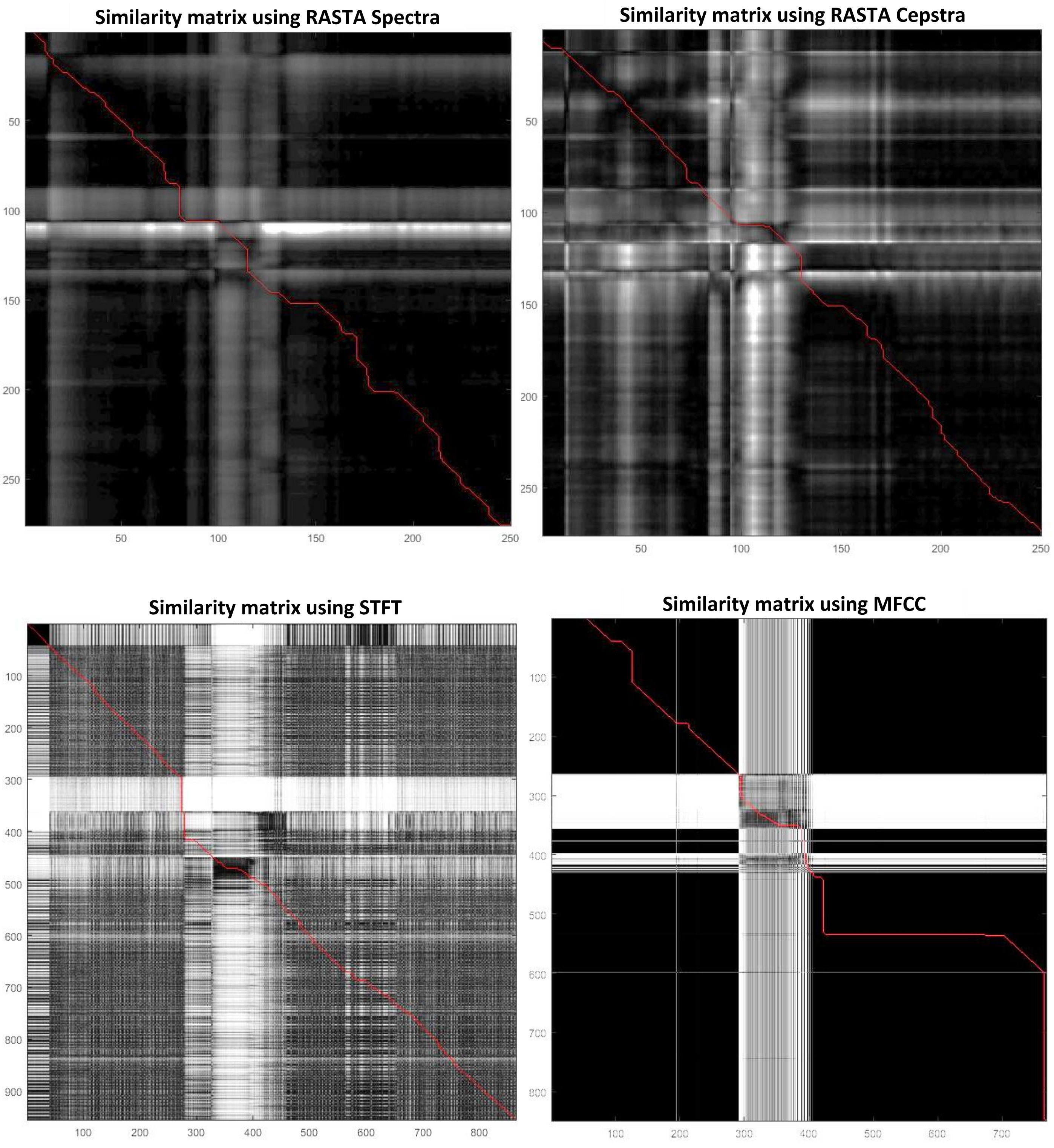
Similarity matrix using MFCC

Figure 18: Similarity matrix by using different techniques between signals ‘satu1’ and ‘satu2’ of same speaker

Figure 18 then shows the similarity matrix between the signals ‘satu1’ and ‘satu2’ by using different techniques which are STFT, RASTA Spectra, RASTA Cepstra and MFCC. We can see that the path of the red line tends to pick darker blocks in each of the similarity matrix since it will maximize the matching performance. Among all the similarity matrix, we can see from the pattern of the red lines in similarity matrix using RASTA methods illustrates that this method has minimum distance between the two signals. This is because the red line is closer to the diagonal indicates that the signals are closely similar. Note that minimum distance is identical to maximum similarity.

4.2.2 Word ‘satu’ and ‘dua’ by the same speaker

Figure 16 shows the similarity matrix between the signals of different words which are signals ‘satu’ and ‘dua’ by using the similar techniques as in figure 17 and 18 which are STFT, RASTA Spectra, RASTA Cepstra and MFCC. These two signals are recorded of different words which are Malay words produced by the same speaker saying ‘satu’ and ‘dua’. These two signals recorded with the same duration of 2 seconds. This figure aims to show that the same length of the signals does not influence the similarity matrix. However, it is based on the similar features or information carried by the signals. As we can see in figure 19, it is clearly shown that there is no any red straight line in any of the similarity matrix indicates that there is no distance match between the signals. Among all the similarity matrix, we can see from the pattern of the red lines in similarity matrix using RASTA methods illustrates that this method has maximum distance. This is because the red line is far from the diagonal indicates that the two signals compared are very different.



Similarity matrix using RASTA Spectra

Similarity matrix using RASTA Cepstra

Similarity matrix using STFT

Similarity matrix using MFCC

Figure 19: Similarity matrix by using different techniques between signals ‘satu’ and ‘dua’ of same speaker

4.3 DTW distance values analysis

Another experiment is conducted to compare the DTW distance between recorded signals in terms of value by using different methods which are STFT, RASTA Spectra, RASTA Cepstra and MFCC. In this experiment, signal 'satu' was compared with itself to determine the distance value and also the signal 'satu' was compared with another different signal which are 'dua', 'tiga', 'empat' and 'sembilan'. The results of the experiment are shown in table 3.

Table 3: DTW distance comparison in terms of value

Methods	**DTW distance value after comparing signals**					
	'satu' & 'satu'	**'satu' & 'dua'**	**'satu' & 'tiga'**	**'satu' & 'empat'**	**'satu' & 'sembilan'**	**Mean**
RASTA Spectra	0	422	408	471	388	338
RASTA Cepstra	0	85	86	81	90	69
STFT	0	5,094	5,009	4,384	6,424	4,182
MFCC	0	3,930	4,638	3,686	4,193	3,289

Table 4: Difference of DTW distance from mean values

Methods	**Difference of DTW distance from mean values**				
	'satu' & 'satu'	**'satu' & 'dua'**	**'satu' & 'tiga'**	**'satu' & 'empat'**	**'satu' & 'sembilan'**
RASTA Spectra	338	84	70	133	50
RASTA Cepstra	69	16	17	12	21
STFT	4,182	912	827	202	2,242
MFCC	3,289	3,930	4,638	3,686	4,193

Based on the table 3, we can see that when similar signals being compared, it shows the distance value of zero which means there is no distance between the two signals and indicates that the two signals are perfectly matched. When different signals were compared, we can see that the value is no more zero and show high values. This indicates that the DTW distance between the signals compared is large which means the signals are not similar. The higher the DTW distance value, the larger the DTW distance between the two signals.

Table 4 then shows the difference of DTW distance values with the mean values for each of the method. From these values, we can see that the values for both RASTA methods are not as high as shown by both STFT and MFCC methods although the signals compared are not identical. Based on the observation, DTW distance between two different words may not be enough to differentiate between those signals. Thus, analysis on Eigenvector distance is conducted.

4.4 Eigenvector distance value between signals

Following results show the comparison of distance between two signals by using different number of eigenvectors. The similarity value that is close to 1 indicates that the two signals are closely similar and vice versa. The similarity value is calculated using the formula shown below:

$$a \cdot b = ||a|| \, ||b|| \cos(\theta) \tag{9}$$

$$\cos(\theta) = \frac{a \cdot b}{||a|| \, ||b||}, [0, 1] \tag{10}$$

In this experiment, there are three different signals by the same speaker saying SATU, DUA and SEMBILAN. For each signal, there are 20 samples which are the different styles of saying SATU, DUA and SEMBILAN by the same person. The following figures show the patterns of similarity values by using different number of eigenvectors for RASTA and STFT methods. The calculation is obtained by taking average of the eigenvector distance from a set of 20 samples of each signal.

Figure 20 shows the patterns of similarity value as we increase the number of eigenvectors of two speech signals extracted using RASTA. The patterns consist of comparison between signals of the same speaker saying SATU, DUA and SEMBILAN. Based on the graph, we can observe that the patterns across all signals are similar where the signals show an increasing similarity from the lowest number of eigenvectors throughout the highest number of eigenvectors. Only few graph shows slightly different pattern from others.

On the other hand, figure 21 shows the patterns of similarity value as we increase the number of eigenvectors of two speech signals extracted using STFT method. Similar to figure 20, the patterns in figure 21 consist of comparison between signals of the same speaker saying SATU, DUA and SEMBILAN. Based on observation, the patterns are almost similar for all the graphs which achieve highest similarity value in the lowest number of eigenvectors but then experience sudden drop and lastly have constant increasing trends as the number of eigenvectors increased.

Based on the two graphs, we can see that RASTA method shows a better result because it has more consistent similarity values compared to STFT method. Most of the energy of the signals are confined by the 3 number of eigenvectors. As we can see at 3 numbers of eigenvectors in both graphs, RASTA shows difference similarity values for the signals compared. For STFT, we can see that all similarity value is high at 3 number of eigenvectors although the signals compared are significantly different. Thus, we can say that RASTA is better than STFT in terms of eigenvector distance.

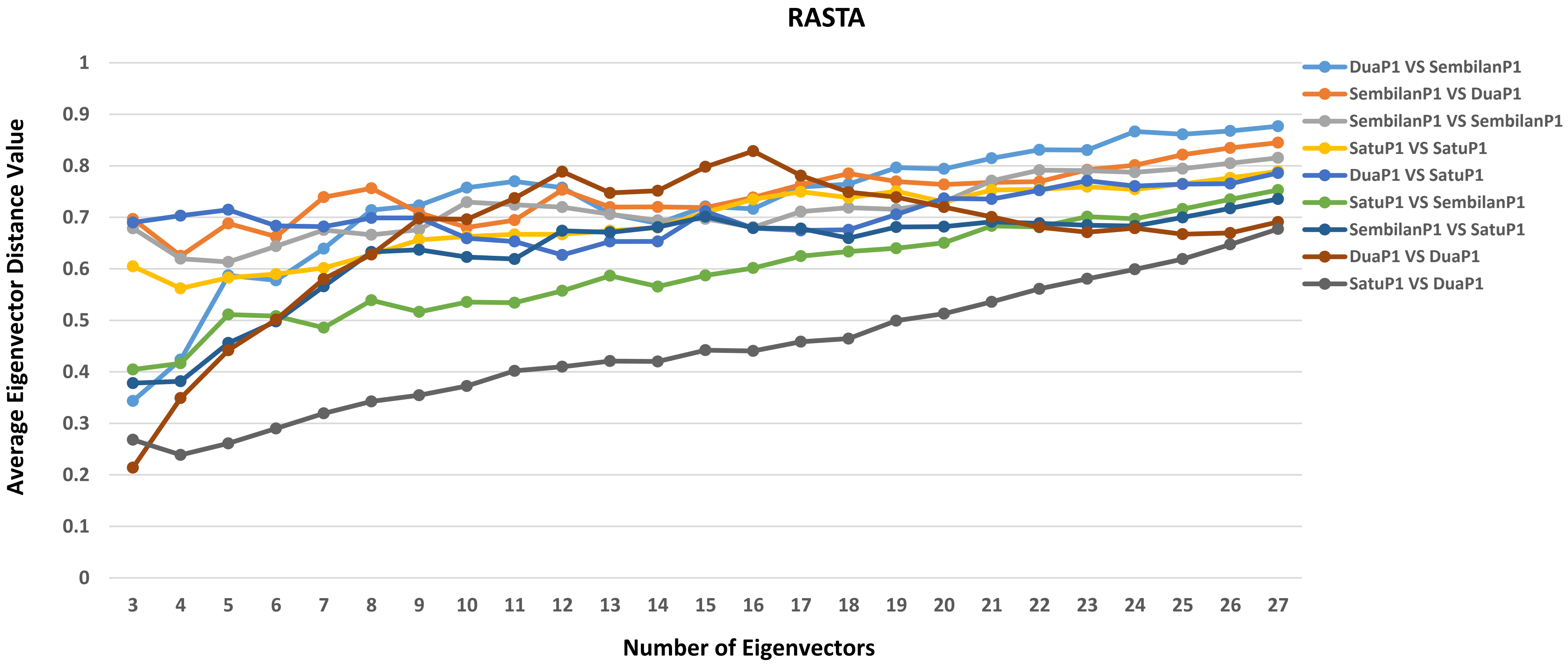


Figure 20: Eigenvectors comparison for RASTA method

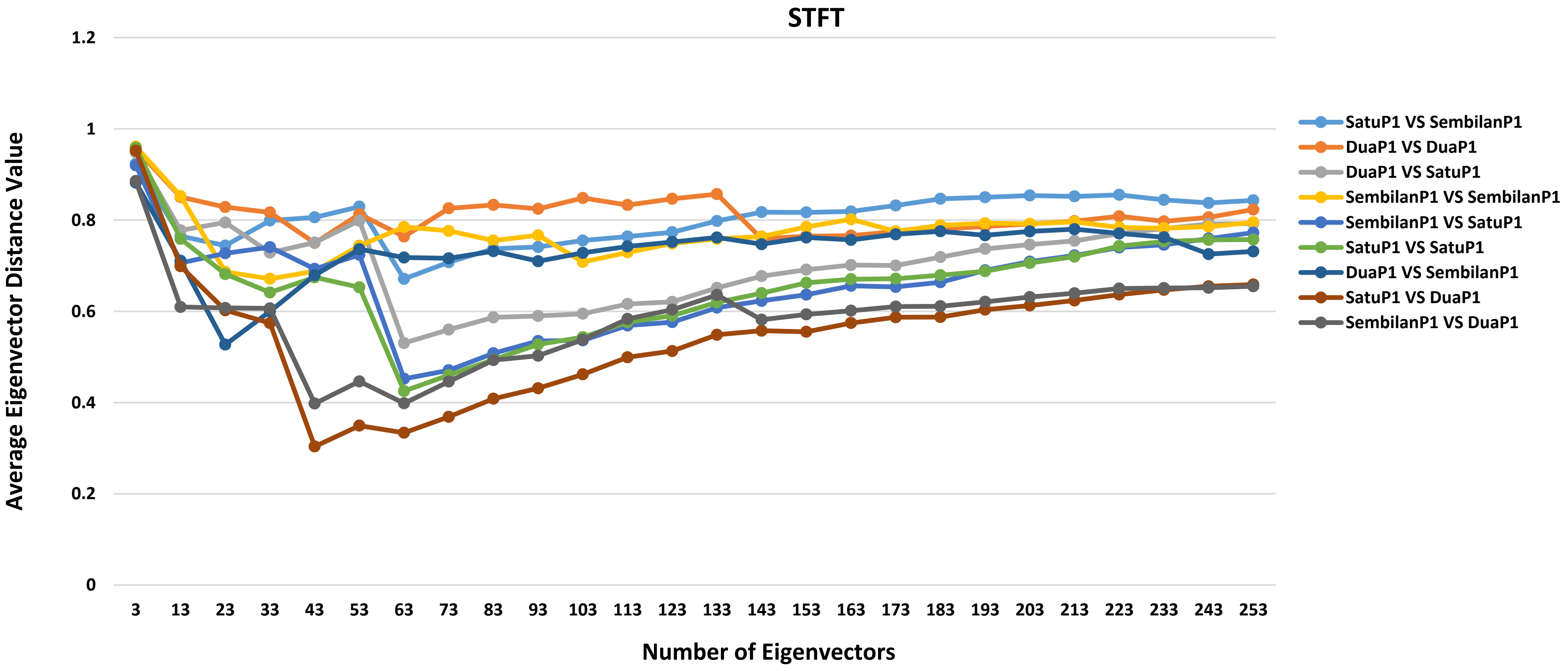


Figure 21: Eigenvectors comparison for STFT methods

CHAPTER 5

CONCLUSION AND RECOMMENDATION

As a conclusion, the techniques of speech recognition system have been shown. Their characteristics have been discussed and analyzed. Based on the experiments conducted, the results obtained give some overview on the performance of certain existed methods such as MFCC, STFT and RASTA for speech recognition. We can know the DTW distance value of compared signals which indicates the similarity between the two signals. Based on the comparison between RASTA and MFCC, we can conclude that RASTA is better in extracting the important features or information from the signal compared to MFCC. This is shown in the spectral features produced by both methods. Experiment on the eigenvector distance value comparison between RASTA and STFT has been conducted and the results show that RASTA is better in terms of its consistency in similarity value as the number of eigenvectors increased. The Singular Value Decomposition (SVD) technique also plays a vital role in this work. This is because more direct results can be seen when eigenvectors are involved as shown in the eigenvector distance value comparison experiment. The comparison on the DTW distance values by using different methods only is not decent enough to differentiate between the two signals. The signals then can be differentiated when eigenvectors are involved. Further experiment need to be conducted to investigate the robustness use of RASTA in differentiating different Malay spoken words.

REFERENCES

[1] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, "A review on speech recognition technique," *International Journal of Computer Applications,* vol. 10, pp. 16-24, 2010.

[2] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara,* vol. 1, p. 67, 2005.

[3] J. T. Garrett, K. W. Heller, L. P. Fowler, P. A. Alberto, L. D. Fredrick, and C. M. O'Rourke, "Using speech recognition software to increase writing fluency for individuals with physical disabilities," *Journal of Special Education Technology,* vol. 26, pp. 25-41, 2011.

[4] T. B. Amin and I. Mahmood, "Speech Recognition using Dynamic Time Warping," in *Advances in Space Technologies, 2008. ICAST 2008. 2nd International Conference on*, 2008, pp. 74-79.

[5] J. Zhang and M. Zhang, "Speech recognition system based improved DTW algorithm," in *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, 2010, pp. 320-323.

[6] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, "Comparative study of automatic speech recognition techniques," *IET Signal Processing,* vol. 7, pp. 25-46, 2013.

[7] S. Singhal and R. K. Dubey, "Automatic speech recognition for connected words using DTW/HMM for English/ Hindi languages," in *2015 Communication, Control and Intelligent Systems (CCIS)*, 2015, pp. 199-203.

[8] S. Jendoubi, B. B. Yaghlane, and A. Martin, "Belief Hidden Markov Model for speech recognition," in *Modeling, Simulation and Applied Optimization (ICMSAO), 2013 5th International Conference on*, 2013, pp. 1-6.

[9] F. Rosdi and R. N. Ainon, "Isolated malay speech recognition using Hidden Markov Models," in *Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on*, 2008, pp. 721-725.

[10] A. Shaukat, H. Ali, and U. Akram, "Automatic Urdu Speech Recognition using Hidden Markov Model," in *Image, Vision and Computing (ICIVC), International Conference on*, 2016, pp. 135-139.

[11] N. Souissi and A. Cherif, "Speech recognition system based on short-term cepstral parameters, feature reduction method and Artificial Neural Networks," in *Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on*, 2016, pp. 667-671.

[12] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, 2016, pp. 493-497.

[13] S. C. Sajjan and C. Vijaya, "Comparison of DTW and HMM for isolated word recognition," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, 2012, pp. 466-470.

[14] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in Speech Recognition System," in *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, 2016, pp. 498-502.

[15] J. Zhang, "Research of improved DTW algorithm in embedded speech recognition system," in *2010 International Conference on Intelligent Control and Information Processing*, 2010.

[16] P. Sadasivan and D. N. Dutt, "SVD based technique for noise reduction in electroencephalographic signals," *Signal Processing,* vol. 55, pp. 179-189, 1996.

[17] R. Rahmat, N. S. Kamel, and N. Yahya, "Principle subspace-based signature verification technique," in *Innovative Technologies in Intelligent Systems and Industrial Applications, 2009. CITISIA 2009*, 2009, pp. 317-321.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing,* vol. 2, pp. 578-589, 1994.

[19] R. Rahmat, N. S. Kamel, and N. Yahya, "Subspace-based signature verification technique using reduced-sensor data glove," in *Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on*, 2009, pp. 83-88.