A NOVEL APPROACH FOR CONTINUOUS SPEECH TRACKING AND DYNAMIC TIME WARPING

Adaptive Framing Based Continuous Speech Similarity Measure and Dynamic Time Warping using Kalman Filter and Dynamic State Model

Wasiq KHAN

Submitted for the Degree of
Doctor of Philosophy

School of Electrical Engineering & Computer Science
University of Bradford

2014

ABSTRACT

Wasiq Khan

A Novel Approach for Continuous Speech Tracking and Dynamic Time Warping

Adaptive Framing Based Continuous Speech Similarity Measure and Dynamic Time Warping using Kalman Filter and Dynamic State Model

Keywords: Speech Tracking, Dynamic Time Warping, Kalman Filter, Dynamic Noise Filtration, Adaptive Framing, Keyword Spotting, Template Matching, Similarity Measurement.

Dynamic speech properties such as time warping, silence removal and background noise interference are the most challenging issues in continuous speech signal matching. Among all of them, the time warped speech signal matching is of great interest and has been a tough challenge for the researchers. An adaptive framing based continuous speech tracking and similarity measurement approach is introduced in this work following a comprehensive research conducted in the diverse areas of speech processing. A dynamic state model is introduced based on system of linear motion equations which models the input (test) speech signal frame as a unidirectional moving object along the template speech signal. The most similar corresponding frame position in the template speech is estimated which is fused with a feature based similarity observation and the noise variances using a Kalman filter. The Kalman filter provides the final estimated frame position in the template speech at current time which is further used for prediction of a new frame size for the next step. In addition, a keyword spotting approach is proposed by introducing wavelet decomposition based dynamic noise filter and combination of beliefs. The Dempster's theory of belief combination is deployed for the first time in relation to keyword spotting task. Performances for both; speech tracking and keyword spotting approaches are evaluated using the statistical metrics and gold standards for the binary classification. Experimental results proved the superiority of the proposed approaches over the existing methods.

DECLARATION

I hereby declare that this submission is my own work and the use of all material from other sources has been properly and fully acknowledged. I also confirm that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature: …………………………………………………………..

DEDICATION

To my parents, my wife, daughter, son, and other family members for their patience, understanding, support, and love.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious, the Most Merciful

All praise is due to Allah for His glorious ability and great power for granting me the health, strength, patience, and ability to complete this doctoral thesis.

I would like to express my deepest appreciation to all those who provided me the possibility to complete my research and this thesis. A special gratitude I give to Prof. Daniel Neagu whose contribution in stimulating suggestions helped me to coordinate my thesis especially in layout, statistical analysis, and performance evaluation methods. I would like special thanks to my supervisor, Dr. Rob Holton for his help, guidance, and support during this research and thesis completion process. I would also like to thank Prof. Ping Jiang and Dr. Pauline Chan for providing their help during my research. Special thanks to Ms. Rona Wilson and Prof. Irfan Awan for being patient with me and providing the administrative needs throughout my research period. Thanks to my colleagues, Dr. Kaya Kuru (University of Central Lancashire) and Dr. Mohammad Bilal (University of California, LA) for their time to proof-read the thesis. Last but not the least; I thank all of my family and friends for their moral and financial support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acronym	Meaning
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CAS	Cosine Similarity Measure
CEv	Combined Evidence
DCT	Discrete Cosine transform
DSM	Dynamic State Model
DSP	Digital Signal Processor
DST	Dempster Shafer Theory
DTW	Dynamic Time Warping
DTWC	Constrained Dynamic Time Warping
ED	Euclidean Distance
FT	Fourier Transform
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
KF	Kalman Filter
KWS	Keyword Spotting
LPC	Linear Prediction Coding
MFCC	Mel Frequency Cepstral Coefficients
MLE	Maximum Likelihood Estimate
MMSE	Minimum Mean Square Estimators
OFDM	Orthogonal Frequency Division Multiplexing
PLP	Perceptual Linear Prediction
PPM	Posterior Probability Measure
PSD	Power Spectral Density
QByE	Query By Example
SFS	Speech Filing System
SNR	Signal To Noise Ratio
STD	Spoken Term Detection
STFT	Short Time Fourier Transforms
SVM	Support Vector Machine
TWCST	Time Warped Continuous Speech Tracking

VAD	Voice Activity Detection
VQ	Vector Quantization
WD	Wavelet Decomposition
WT	Wavelet Transform
YAAPT	Yet Another Approach for Pitch Tracking
ZCR	Zero Cross Rate

1. INTRODUCTION

There is a proliferation in the process of Automatic Speech Recognition (ASR) research and development (Li et al. 2014; Baker et al. 2009) in the timeframe for the speech analysis and synthesis systems development. Early days ASR systems were only able to respond a significantly limited set of sound samples that have been improved to sophisticated systems that respond to fluently spoken natural language taking into account the dynamic properties of the spoken language. Despite the fact that literature consists of a variety of methodologies that have been introduced for ASR modelling, there is a limited research work available to deal with time warped continuous speech signal matching. In exception to Dynamic Time Warping (DTW) and some of its enhanced versions, most of the related research is focused on statistical models. Stochastic language modelling based ASR systems are able to deal with large vocabulary and continuous speech recognition (Arora and Singh 2012; Rabiner and Juang 2004). Hidden Markov Model (HMM) based ASR systems are the best examples of such systems that are considered as state of the art. On the other hand, template based approach to speech recognition provided several simple and interesting techniques for speech recognition.

In the template based ASR model, recognition is performed by matching the test word (utterance) with stored template of words and calculating the matching score based on acoustic features (Cheng et al. 2014; Arora and Singh 2012; Rabiner and Juang 1993). The DTW and vector quantization (VQ) based ASR is the best examples of such systems. In relation to continuous speech signal matching, there are some limitations associated to all aforementioned techniques. For example, statistical ASR approach depends upon the phonemes probability assignment which if not assigned correctly, leads to mis-recognition. The pattern based methods are limited to specific vocabulary of words with high computational cost for system training. The boundary constraints and pruning of search space degrades the similarity matching performance in DTW approach. Similarly, static frame size and a single source of information for decision making in DTW approach are the concerning issues (Cheng et al. 2014; Ratanamahatana and Keogh 2005).

The proposed research work emphases on a Time Warped Continuous Speech Tracking (TWCST) and similarity measurement based on the acoustic features in the speech utterance. Speech is a rich information carrying signal consisting of useful features that provide significant knowledge to represent a speech utterance. Therefore, in present work, detailed review in the area of speech signal enhancement, spectral analysis, and feature extraction is conducted. The existing silence removal approaches are reviewed and a new approach is introduced that is based on combination of simultaneous information from time and frequency domain features. In addition, a recursive feedback system is deployed that uses the Kalman Filter (KF) for tracking the current position of test speech frame with respect to template speech. Usage of a KF provides novelty in terms of adaptive frame size and fusion of multi-source information for decision making that resolve the major issues related to the existing template matching based similarity measurement approaches. Also, a new approach for keyword spotting is introduced that uses the Dempster’s theory of combination of evidence from multiple resources of information. The usage of theory of evidence make the existing keyword spotting approaches more reliable and trustable in terms of decision making based on multi information resources.

1.1. Motivation

Similarity measurement between speech signals aims at calculating the degree of similarity using the acoustic features. Nowadays, it is receiving much interest due to the large volume of multimedia information. A time warped speech signal matching is related to pattern based speech recognition. However, there is a limited effort available in the literature that deals with continuous speech signal matching and time warping issues. Most of the available research work for speech signals matching is limited to isolated word matching and keyword spotting that is based on statistical modelling, acoustic features, and DTW. In relation to continuous speech signal matching, there are various limitations associated to all aforementioned techniques. For example, statistical ASR approach is limited to a specified language model as it involves the language grammar and a learning process that uses the transcribed data. Moreover, it depends upon the phonemes probability assignment which if not assigned correctly, leads to mis-recognition. Pattern based methods are limited to specific vocabulary of words with high computational cost for system training. Also, there is a trade-off between recognition efficiency and vocabulary size.

The DTW is a popular technique that has been used for time warped signal alignment. However, there are several drawbacks associated with the DTW approach as described below that are critical to be resolved in the proposed research work. To trade-off between the search space pruning and similarity performance is a challenging issue associated with the DTW approach. In addition to this, the DTW uses fixed frame size in speech signal matching that lakes the use of global time warping in terms of continuous speech signals. However, it would be much better to use dynamic frame size that can be adapted recursively in a similar fashion as of the natural time warping phenomenon of a speech signal. In addition, DTW measures the warping distance based on a single source of information that uses traditional distance metrics (i.e. Euclidean distance, cosine similarity). In contrast to this, the use of a dynamic model that considers the noise factors and provides an additional support to the distance metric base warping distance calculation. A TWCST approach having the aforementioned capabilities would be a novel as introduced in the current research.

1.2. Research Questions

Following from the motivations listed in the section above, one may find the following challenges to be addressed in this work:

1) *Is there a similarity measure that is able to estimate the corresponding positions for two time-warped continuous speech signals consisting same speech content spoken by the same speaker at different time?*

2) *Is there a reliable approach available that can produce template matching based speaker dependent keyword spotting without using the transcribed data?*

To answer the above research questions, it is necessary to consider the challenges associated with a DTW approach as discussed in (Cheng et al. 2014; Abad et al. 2013; Ratanamahatana and Keogh 2005). A speech tracking approach will be helpful for proof reading and for children’s speaking and reading skills improvement. Similarly, it would be helpful to provide a feedback to a user who wants to learn the speech contents by heart. Likewise, the keyword spotting in continuous speech is an interesting research topic with a number of useful real time applications. For example, search by association where a user has no specific aim other than finding interesting things. It can be used to search for a specific word in a long speech recording. Moreover, keyword spotting applications may be helpful for human-computer interaction where a user can pass the command by speaking out a specific word and machine performs relevant task accordingly. In a similar way, keyword spotting may be very useful for intelligent agencies to keep track of specific target word in trapped calls.

1.3. Research Methodology

Several related formal analytical approaches and research methodologies were studied to establish a research design for the development of an appropriate experimental system. A quantitative research method is used to address the critical questions which are the formal, objective, and systematic processes to: (1) describe and test relationships, and (2) examine cause and effect interactions among variables (Burns and Grove 1993).

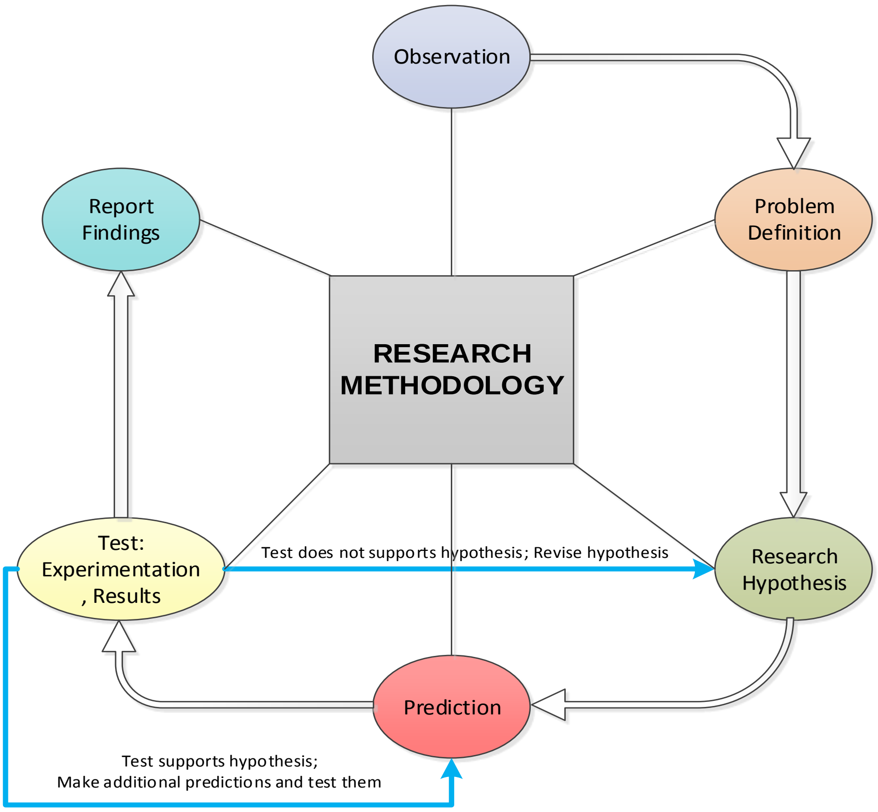

Observation
Report Findings
Problem Definition
RESEARCH METHODOLOGY
Test: Experimentation , Results
Test does not supports hypothesis; Revise hypothesis
Research Hypothesis
Prediction
Test supports hypothesis;
Make additional predictions and test them

Figure 1-1: Research Methodology for the Current Research

The quantitative research design has the advantage of finalizing the results and proving/disproving the hypothesis and known as a standard experimental method of the scientific disciplines (Shuttleworth 2008). Overall structure of the quantitative research approach for this work is presented in the Figure 1-1.

1.3.1. Aims and Objectives

The main **aim** of this thesis is twofold that is:

1) To propose a speaker dependent speech tracking approach that deals with the dynamic time warping to produce a similarity calculation between two time-warped continuous speech signals.

2) To propose a reliable speaker dependent keyword spotting approach that is able to find a target keyword in a continuous speech signal.

In order to achieve the goal, following **objectives** have been set:

A. To amalgamate a number of methodologies to measure the similarity between two time warped continuous speech signals, that includes:

- **a.** Speech enhancement.
- **b.** Speech segmentation.
- **c.** Spectral analysis.
- **d.** Time and frequency domain feature extraction.
- **e.** Dynamic state model for segmented speech.
- **f.** Similarity measurement using different similarity measures.
- **g.** Kalman filter based feedback system for the recursive frame position estimation in template speech.

B. To further improve the existent isolated word matching and keyword spotting techniques at a level where they are able to identify an utterance/keyword in a continuous speech signal on the basis of acoustic features using the Dempster's rule of mass combination.

C. To design a framework that sequentially combines aforementioned techniques to provide a recursive frame size adaptation along the time progression.

D. To design a dynamic filter based on wavelet decomposition that is capable of removing the unnecessary time and frequency components from the speech signal.

E. Deployment of the posterior probability measure for the purpose of speech signal matching in the presence of background noise.

F. To Critically evaluate and test the proposed approaches for speech tracking and keyword spotting using the performance evaluation and validation methods of a binary classifier and comparing with the existing state-of-the-art techniques.

1.3.2. Data Collection

For the proposed research study, data is collected by various open source and proprietary speech dataset that consists of speech recorded by speakers from diverse backgrounds, age groups, gender, and speaking accents. The speech corpuses include American Rhetoric's (Michael 2001), CMU ARCTIC (Festvox, arctic database), (Online audio stories), Mobio (McCool et al. 2012), and Wolf (Hung and Chittranjan 2010) are requested from IDIAP research institute. These corpuses contain speech phrases for isolated words, short phrases, paragraphs contents, and long speech recordings. In addition, a case study is conducted for the proposed research studies and a dataset is recorded by 30 speakers from diverse background and speaking accent. This data is naturally time warped as it recorded at different time, hence each sub-word in a speech recording is naturally time warped. A detailed description of the speech corpus used for experimental analysis is provided in Section 4-4 and Table 4-3.

1.3.3. Method of Data Analysis and Processing

Speech processing is a research area that consists of a wide range of working platforms available in the market. There are several platforms that can be used to analyse the speech signal depending upon the required task. In the proposed research work, Audacity is used for recording the new data. It provides a high range of processing tools that are very easy to understand and use. Another most common platform known as Speech Filing System (SFS) is used in the current research work. The main purpose of this tool is to analyse the speech feature in time-frequency domain. An SFS platform provides deep level information about a speech segment in terms of its properties. Similarly, PRAAT is most feasible and commonly used platform for the speech labelling and annotations. Matlab is a powerful tool that is used for analysing the speech data and results. Detailed description of these toolboxes is provided in Appendix C.

Although the aforementioned platforms provide a variety of speech processing functionalities, yet Matlab is a well-known toolbox for signal processing. It consists of a wide range of built-in routines for speech processing. To conduct the proposed research study, Matlab 2009a is used for most of the speech processing tasks and the development of a demo work. Small routines are utilized for different tasks that include silence removal, noise reduction, sample rate conversion, framing, spectral analysis, feature extraction, and distance metrics etc. In addition, it provides a variety of data presentation tools that can demonstrate the output results in form of two and three dimensional representations. A detailed description for each toolbox is provided in Appendix C.

1.3.4. Experimental Methodology

The proposed research contributions for continuous speech tracking and keyword spotting are based on multiple processes that run sequentially to perform the desired task. Each component (process) is related to a different research area that is explored individually. In the proposed research work, these components include speech enhancement (sampling, noise reduction and silence removal), framing, spectral analysis and feature extraction, distance metrics, dynamic state modelling, belief combination, and feedback systems. To conduct the research in the aforementioned areas of the proposed research contributions, following methods are deployed:

A). Literature Review

A comprehensive literature review is conducted for related research areas that provide the information of latest research work associated to each sequential component of the TWCST and keyword spotting approaches. Speech enhancement, spectral analysis, and feature extraction related research is discussed in Chapter 2. Existing research work related to distance metrics, dynamic time warping, keyword spotting, query bay example, and feedback systems are presented in Chapter 3.

B). Mathematical Formulation

To achieve the research objectives of TWCST approach, a mathematical model is presented that describes the whole system in a set of equations. Chapter 4 presents a mathematical model for the proposed research contribution for time warped continuous speech tracking. Wavelet decomposition based TWCST approach is presented in Chapter 5. To address the second research question, the mathematical formulation of keyword spotting approach is presented in Chapter 6.

C). Theoretical justifications

To empower the research contribution (research question 1), a detailed discussion is presented in Section 5.4 that compares the key advantages of the contributed work in terms of TWCST approaches with the existing techniques. A justification for each contribution is discussed that empowers the theoretical aspects of the proposed research work.

D). Descriptive Design (Case Studies)

To prove whether a scientific theory and model work in real world, a case study research design is also useful. As the idea of TWCST is introduced very first time as a part of current research study, the case studies empower the validity of hypothesis in the real world. For the aforementioned research questions, case studies are conducted with an average and diverse population while performance is evaluated using the statistical analysis. Details of the case study dataset, population, environment, and statistical results are presented in performance evaluation sections of keyword spotting and continuous speech tracking.

E). Validity and Performance Evaluation

The performance is evaluated using true experimental design which has the advantage of evaluating the results by statistical analysis. The overall model is tested on the collected speech data and the performance is validated using binary classification validation methods. The criteria for performance evaluation are set according to the gold standards used for validation of a binary classifier (i.e. sensitivity, specificity, likelihood ratios, absolute error rate, and F1 Score) which are described in Chapter 4, section 4.4. The performance is compared with existing techniques to measure the degree of validity of the research hypothesis. A detailed discussion on performance evaluation is presented in both scenarios in Chapter 4 and Chapter 5. The performance evaluation for research question 2 is presented in Chapter 6.

F). Report the Findings

Finally, the research contributions are published in peer reviewed journals and conference papers as described in the contributions list of Appendix A.

1.3.5. Ethical Considerations

Shamoo & Resnik (2003) list several ethical principles such as honesty, objectivity, integrity, openness, respect for Intellectual Property (IP), confidentiality, responsible publication, responsible mentoring, social responsibility, non-discrimination, competence, legality, and human subjects protection. In particular, regarding human subjects, there is a need to conduct research with a view to minimising risks and maximising advantages for the subjects while respecting their privacy and maintaining confidentiality. As we planned to conduct research on speech tracking and keyword spotting in multiple speech recordings and investigate the effects of time warping due to speaking speed variations, there was a need to engage with human subjects for speech data collection. This necessitated ethical compliance of the study based on a clearly specified framework of practice.

Informed consent: Subjects have been orally informed of the procedures in the research study prior to being asked for their consent.

Confidentiality: All data that is recorded personally is rendered anonymous and kept confidential at all times. Once the study will be completed, the personally collected data would be kept secret. Identifying information would be kept strictly confidential and accessible only to those directly involved in the study. Only essential personal information would be sought from the participants in the study and all such information would be anonymised.

1.4. Research Contributions

The achievements and contributions in the proposed research work that are presented in this thesis are listed as follows:

A. Dynamic State Model for Speech Tracking

In the speech processing related literature, to the best of our knowledge, there is no clue of time warped speech tracking and similarity measure that may be a great interest for human-machine interaction applications. The existing speech signal similarity matching techniques have been reviewed in order to achieve the uniqueness of the proposed approach for the continuous speech tracking. Based on equations for object's linear motion, a Dynamic State Model (DSM) is presented (Chapter 4, Section 4.3) that considers the input or test speech signal as a frame by frame linearly moving object along the template/reference speech signal with the progression of time (Khan and Holton 2015; Khan et al. 2014).

B. Dynamic Time Warping and Frame Size Adaptation

In the proposed approach for TWCST, a novel idea of adaptive frame size is introduced (Chapter 4, Section 4.3.6) that changes the template speech frame size dynamically at each time step with respect to input speech. Initially; the template frame size is kept same as of test frame but it dynamically changes at each time step with respect to the speaker's speed of input (test) speech. This implies that rather than finding the overall warped distance as in traditional DTW, the warped distance can be minimised by recursively and dynamically adapting the template frame size according to input speech, on the basis of position estimated by dynamic model at previous step (Khan and Holton 2015; Khan et al. 2014).

C. Kalman Filter and Position Estimation in Continuous Speech Tracking

In the proposed approach for speech tracking, a Kalman filter (KF) is used for the input speech frame position estimation with respect to template speech signal. The two position observations; from DSM and similarity measure are forwarded to KF along with process and measurement noise covariance. The KF process these inputs and provides a best position estimate in the template signal corresponding to test speech frame at current time. This position estimate is further processed by adaptive framing process to predict the new template frame size for next time step. The whole feedback cycle runs recursively until the end of test or template speech signal. The best estimated positions in the template speech relative to input speech frames are recorded along with a similarity score (Khan and Holton 2015; Khan et al. 2014).

D. Dynamic Filtration of Speech Signal Using Wavelet Transform

An efficient technique is presented in the proposed research work for the Wavelet Decomposition (WD) based time and frequency band filtration (Khan et al. 2014). The filtration process improves the matching performance because of the unnecessary components filtration. In the literature, Short Time Fourier Transform (STFT) has been successfully used for the frequency domain representation of speech signal. The disadvantage of the STFT is the low time-frequency resolution as compared to WD that provides simultaneous representation of time and frequency of speech signal (Fugal 2009). A major advantage of three dimensional representations by WD is the filtration of unnecessary segments and frequency bands (levels) from the speech signal that improves the test and template matching outcome (Chapter 5).

E. Keyword Spotting Based on Acoustic Features and Theory of Mass Combination

In the proposed research work, a variety of time and frequency domain acoustic features are analysed and performance of the proposed technique (Khan and Holton 2015) is compared with existing word identification approaches. The proposed wavelet based dynamic filtration is applied to input speech and then passed for the feature extraction process. The extracted features are forwarded to multiple similarity measure algorithms that calculate the matching scores for the target speech utterance and all corresponding frames of template speech. These scores are forwarded to a Decision Support System (DSS) that uses the Dempster's rule of mass combination. The Dempster's rule considers these scores as observer's beliefs with the predefined weights and measures a combined belief. All frames in the template speech crossing a pre-set threshold value for the combined belief are identified as spotted keywords (Chapter 6).

F. Deployment of Posterior Probability Measure for Speech Signal Matching

An interesting technique for image blob matching known as Posterior Probability Measure (PPM) is deployed for the first time in the proposed research work (Chapter 6, Section 6.4). Originally, the PPM was proposed by (Fing et al. 2008) for blob matching and tracking the target object in the image. The PPM has a unique property of differentiating the background noise components from speech components when measuring the similarity between target and reference speech models. The separation of these features shared by both target and background models produced robust results in object tracking and localisation. It results more reliable and tolerant pattern match to varying model scale. Experiments proved the superiority of PPM over the existing similarity measure techniques; more specifically in the presence of background noise. The minimization of noise effects on similarity measure performance is very useful that is achieved in the proposed research (Khan et al. 2012) with the deployment of PPM in speech signal area to degrade the background noise influence from a speech signal.

G. Vector Addition Based Similarity Measure

A literature review is conducted for the comparison of various similarity measurement algorithms for isolated speech utterance matching. In addition, by comparing the mathematical implementation of the most commonly used cosine similarity measure and Euclidean distance, a new similarity measure (Resultant Vector) is proposed in the current research work (Khan et al. 2013) that uses the head to tail rule for vector addition. Performance of the newly introduced similarity measure is compared with Euclidean distance, vector cosine angle distance, and Bhattacharyya coefficients that show satisfactory results in terms of matching output (Chapter 6, Section 6.5).

1.5. Report Outline

The research work and contribution presented in this thesis are based on diverse areas that include speech enhancement, distance metrics, feedback system, and decision support systems. Therefore the thesis chapters are organised according to the area of studies. This thesis consists of seven chapters. Chapter 1 introduces the research area of time warped continuous speech tracking and keyword spotting together with the speech recognition concepts employed in current research work.

The scope, motivation, research questions, aims, objectives, research methodology, and outcomes of the thesis have been highlighted in this chapter. Chapter 2 presents the background concepts and literature review of the speech processing used in the current research study. Speech enhancement, segmentation, spectral analysis, and feature extraction are the target modules that are reviewed corresponding to the speech tracking and similarity measurement.

Distance metrics and Kalman filter's theoretical concepts and research work in the related area of speech processing is presented in chapter 3. Chapter 4 presents the major contributions of proposed research work towards a TWCST and similarity measure approach. Mathematical formulation and the performance analysis are presented in this chapter. The Sequential process is presented in a flowchart. The overall procedure is further divided into four subsections detailing the design of speech tracking structure along with dynamic state model and KF deployment. First two subsections are based on speech enhancement and feature extraction. Third section provides the detailed formulation of a dynamic static model that considers the speech signal as a linearly moving object with uniform velocity. Finally, the last section presents the formulation of KF for speech tracking approach and producing a recursive position estimate that is used for dynamic frame size measurement. Chapter 5 addresses the contributed work towards an alternative approach for TWCST that uses the concept of dynamic noise filtration. A dynamic noise filter is introduced that uses the WD to represent the speech signal in a 3 dimensional forms. Rest of the model for this approach is same as discussed in Chapter 4. The performance is evaluated by statistical analysis methods used for the validation test of binary classification.

Chapter 6 demonstrates the contributions towards the keyword spotting in continuous speech as well as for the isolated word matching in the presence of background noise. In addition, a comparison between the existing similarity measures is presented in this chapter. The keyword spotting locates multiple occurrences of a target utterance in a long continuous speech recording. A dynamic noise filtration method using WD is introduced. A novel idea of deployment of Dempster- Shafer's theory of evidence is presented for the key word spotting approach. The performance is compared with the existing keyword spotting methods and the statistics are presented in the performance evaluation section. Chapter 7 aims to provide a conclusion of the thesis and relates the main objectives to the contributions made in the current research work. In addition, it provides detailed recommendations for the possible future directions. Finally, Appendices present the publications list, technical reports, detailed statistical results for TWCST approaches; Matlab scripts for the contributed work, and details about processing and simulation tools.

2. SPEECH PROCESSING

2.1. Scope

To achieve the specified research objectives towards the TWCST and keyword spotting approaches, it is necessary that the query and reference speech signals are pre-processed and enhanced. This enhancement can be made in terms of silence removal and background noise reduction while considering the speech intelligibility. Enhanced speech signal is then forwarded for the spectral analysis and feature extraction process. This chapter consists of three sections which aim to provide a comprehensive insight into the speech signal processing and current level of research performed in this area. A comprehensive review is presented in terms of different aspects of speech processing that includes: (1) speech enhancement, (2) segmentation (i.e. windowing and framing), and (3) spectral analysis. Detailed implementation of Fast Fourier Transform (FFT), STFT, and Wavelet Transform (WT) is presented along with their advantages, disadvantages, and applications in the related area. The last section is based on a comprehensive discussion of how different features are extracted in time and frequency domain that possess enough information for the identification and similarity matching of speech utterances.

2.2. Speech Signal Processing

“A digital signal processor is an integrated circuit designed for high speed data manipulations and is used in audio, communications, image manipulation, and other data acquisition and data control applications” (Digital Signal Processing, Pp. 1-18). Speech processing is the application of digital signal processing (DSP) to process and analyse the speech signals. Speech signal consists of several kinds of information that may be useful in real life applications. As an example; speech signal carries the information about meaning of the contents a speaker wishes to impart, speaker information, emotions in the speech, and much more. This information is useful for development of the applications in diverse areas in real life. To extract such information, speech signal can be processed either in time domain or frequency domain. A time-domain representation of speech signal shows how the signal changes over time. Speech signal can be converted from time domain to the frequency domain by a transformation process. A number of transforms are available that are explained later in this chapter. The most common purpose of speech signal analysis in frequency domain is to extract those features that can’t be retrieved in time domain. These features may help for the identification of speech utterance on the basis of their dominant acoustic properties. Speech processing is a broad and multi-functional term that can be mainly divided into different subtasks that are described in the following sections.

2.2.1. Speech Enhancement

Speech enhancement has been an interesting research area mainly focusing on the suppression of additive background noise (Ephraim and Malah 1985). The term speech enhancement basically refers to the speech quality improvement in terms of degrading all those factors that affect its intelligibility. There are different aspects to be considered for the speech signal enhancement. For example, different kind of noise including additive acoustic noise, acoustic reverberation, and convoluted channel effects, have been studied in the literature using signal resampling, echo suppression and silence removal (Speech Enhancement, Pp. 48-52). Additive noise and convolution effects degrade the intelligibility and ability to listen the contents of speech. Generally, there are multiple aims of speech enhancement such as effective encoding for storage and transmission, improved performance of ASR systems, and acceptability to the human listeners (Sahoo and Patra 2014), (Brookes et al. 2012). In the literature, most of the research work conducted in the area of speech enhancement is related to ASR and speaker recognition system. The most common approaches used for speech enhancement are discussed below.

2.2.1.1. Frame based processing

Speech enhancement methods operate in both; time and frequency domain. For the frequency domain signal enhancement, the sampled input signal $x(n)$ is required to be decomposed into overlapping frames as shown in Figure 2-1. Each frame consists of '*N'* samples and is given by:

$$x(n,l) = w(n)x(n + l(N - M)) \qquad \text{2-1}$$

Where; n = 0...N - 1, 'l' is the frame counter and '*M'* represents the number of overlapping samples with *M*<*N* and represents the time increment between two consecutive frames (number of samples). The $w(n)$ represents the window function that is zero valued out of some chosen interval.

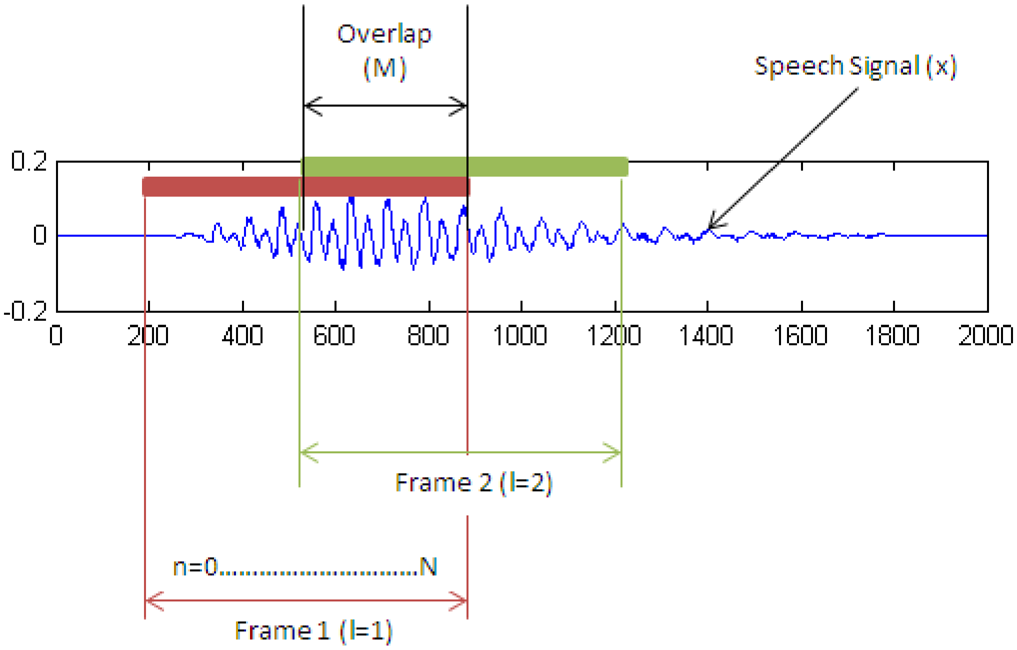


Figure 2-1: Speech Framing and Overlapping

The most common types of window function include rectangular window, hamming window and hanning window. There is a trade-off between time and frequency resolution because of their inverse relationship. Hence, the window length, '*N'* has to be compromised and is typically chosen in the range 10-50 milliseconds for speech signals processing (Ravindran et al. 2010). The original speech signal can be reconstructed if no processing is done on the frame in time or frequency domain. However, distortion artefacts may be introduced due to signal discontinuities at frame edges and aliasing of rapidly changing frequency components in the result of frequency-domain processing. To deal with the problems caused by frequency domain processing and to reconstruct the original signal, windowing function and overlap ratio are used. To provide the perfection in reconstructing the original signal and to deal with discontinuities at the boundaries of frame; a square root hanning window for both; analysis and synthesis of speech signal with 50% overlap is introduced by (Martin and Cox 1999). Similarly, a detailed research analysis is presented by (Allen 1977) and (Allen and Rabiner 1977) to overcome the aforementioned challenges, where a hamming analysis window with three-quarters of overlap is introduced.

2.2.1.2. Sampling

Speech signal is analogue signal that needs to be digitized by a specific number of samples per unit time for further processing. This process is called *sampling* and the value of the signal we choose to retain is determined by sampling rate or sampling frequency that represents the number of samples per second. The amount of time between two successive samples is known as *sampling period* or *sampling interval.* Sampling frequency depends upon the signal's maximum frequency. Harry Nyquist presented his findings about the relation between sampling frequency and signal maximum frequency that is known as *Nyquist sampling theorem*. It states that "a signal must be sampled at a sampling rate equal to at least twice its highest frequency component" (Deng and Shoughnessy 2003). i.e., $f_s = 2f$, where f_s is the sampling frequency and '*f'* represents the maximum frequency component.

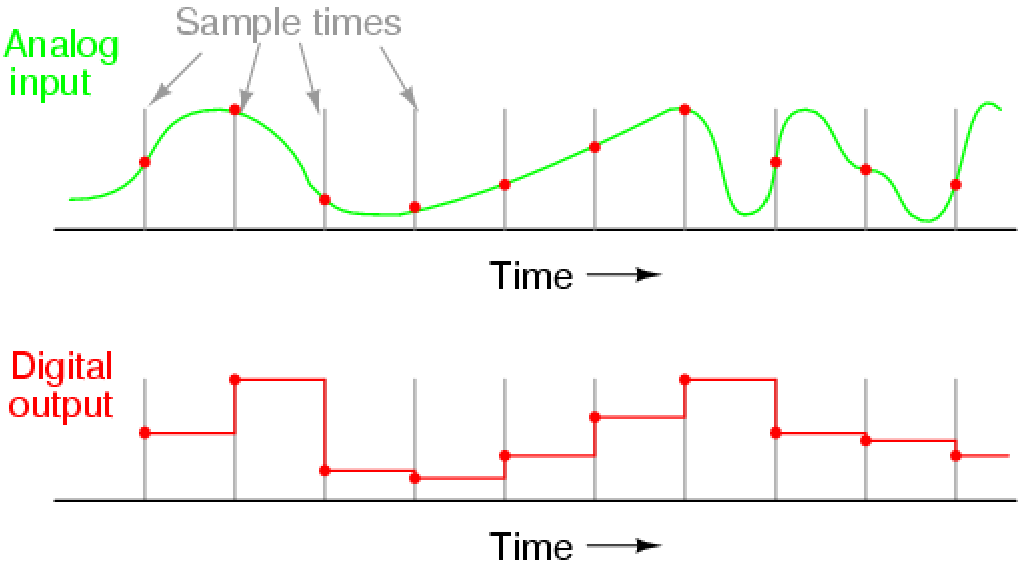


Figure 2-2: Analogue and Digital Signal Representation (Kuphaldt 2007)

The green line in Figure 2-2 represents the continuous analogue signal, whereas red dots indicate samples. The problem of aliasing occurs when a speech signal is sampled under the Nyquist sampling frequency. This means that the sample rate is too low for the higher frequency components of a signal. A best example of aliasing is shown in Figure 2-3. The analogue signal is presented by a solid curve whereas; dotted curve represents the sampled signal formed by circled samples. It is clear that the dotted curve consists of only two cycles instead of 20 cycles of original signal. This is the most common reason of digital speech signal distortion.

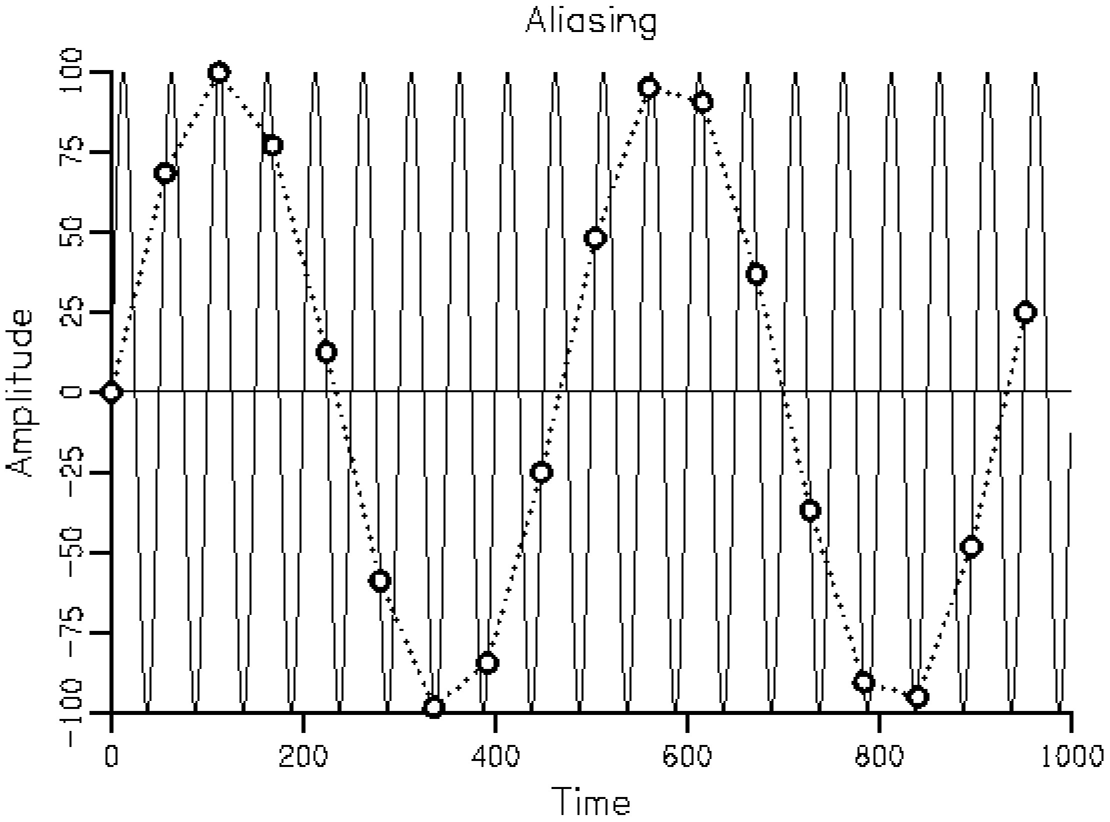


Figure 2-3: Aliasing Due to Low Sampling Frequency (Lieberman and Blumstien 1988)

The aliasing effect can be avoided by pre-filtering the signal so that maximum frequency (highest frequency) of the signal is less than half of the sample rate (Wang 2006). This means that a low-pass filter can be use that cut-off the frequency at half of sample rate. A research is conducted in (Hirsch et al. 2001) for the comparison of speech recogniser performance using different sampling frequencies for a speech signal. It is analysed that word error rate remains almost constant for the 8, 11 and 16 KHz speech signals when tested by HMM based speech recognition system. This is because, human speech is covered between the range of 80 Hz to 3.3 KHz that needs maximum 8 KHz sampling frequency to satisfy Nyquist theorem and to avoid aliasing effects. In the proposed research, a sample rate of 8 KHz is used for the speech signal analysis that is supported by conducted research experimentation as well as by the literature (Lakshmikanth et al. 2014), (Rodman 2006), and (Lieberman and Blumstien 1988).

2.2.1.3. Noise Filtration

Noise filtration is one of the most important processes for the speech enhancement. The main intension is to lower the noise level without affecting the speech quality. Basically, background noise estimation or degradation is needed to increase the speech intelligibility and clarity that can be achieved through applying different speech filtering techniques. A filter is based on electronic circuit that attenuates the unwanted and noisy components from the original signal. There is always a trade-off between the noise reduction quantity and speech distortion that is considered a common challenge for noise filtration process. More noise reduction cause the degradation of speech signal in terms of information lost.

A comprehensive study on speech signal noise filtration techniques is presented by (Lakshmikanth et al. 2014) and (Maher et al. 1992). Different adaptive filtration methods including Wiener filter, line enhancer, multiple microphone, and spectral subtraction were operated on the speech spectrum. The speech signal was segmented over 10 milliseconds duration and FFT is performed with 75% overlap. The sampling frequency was decreased to 8 KHz from 20 KHz and Gaussian distributed noise was generated at 5, 10 and 20dB SNR. The noise reduction performance showed that the adaptive Wiener filter outperforms all other techniques. However, using SNR measures, a spectral subtraction technique provided up to 10 dB improvements in SNR, compared with 6 dB for an adaptive line enhancement method. Spectral subtraction is also a commonly used technique for noise reduction. A speech enhancement approach is presented by (Djigan et al. 1999) for the reduction of additive stationary and quasi stationary colour broad band noise without voice activity detection deployment.

Research conducted by (Chen et al. 2006) evaluates the performance of a Wiener filter for the noise reduction. It is analysed that if linear prediction coefficients of desired clean signal are known; then the clean signal can be achieved with a minor level of speech distortion using these coefficients. In case of no prior knowledge, they introduced a sub-optimal filter on the basis of a free parameter to control the compromise between speech distortion and the level of noise to be reduced. The performance of sub-optimal filter showed that using compromise parameter value 0.7, the noise reduction performance decreased by 10% with a 50% less speech distortion as compared to traditional Wiener filter. In the current research study, a dynamic noise filtration technique is introduced (Chapter 5) based on wavelet based processing that is used for the background noise reduction (Khan et al. 2014).

2.2.1.4. Silence Removal

Speech signal changes its properties with respect to the analysis window or frame size. For a short frame size, speech signal shows very slow variations along the time progression, whereas for long time interval, speech features are dynamic and non-stationary. Normally, speech signal is classified into voiced, unvoiced, and silence components. In a voiced speech, vocal tract produces periodic vibrations by which the fundamental frequency is calculated. All vowels, semivowels, and some consonants like /m, /n/ (i.e. nasal cavity usage) fall in this category. In unvoiced speech on the other hand, the vocal cord does not vibrate, hence the speech produced is much like random nature. The excitation of the vocal tract by a steady air flow creates a turbulent in the region of a constriction in the vocal tract that produces fricatives like /f/, /s/, and /sh/. In the silence part of speech, there is no sound produced resulting very low energy for the signal.

Silence removal plays a significant role in speech processing and recognition applications (Sahoo and Patra 2014). Once, silence segments are identified, these are removed from the speech signal which improves the performance in terms of speech recognition, signal matching and computational cost. The higher frequency of unvoiced speech segments means higher zero crossing rates (Rabiner and Juang 1993). Similarly, silence part can be easily separated by considering low energy segments of speech. On the basis of aforementioned time dependant features of speech signal, an extensive research work has been conducted in the literature for the speech labelling and silence removal. Generally, Zero Cross Rate (ZCR) for silence speech part is always lower than unvoiced speech and higher than voiced speech (Sahoo and Patra 2014), (Sarma and Venugopal 1978). On the other hand, energy value is comparatively higher for voiced speech than the unvoiced and silence segments. In addition to ZCR and energy estimate, Fundamental frequency (F_0) and pitch have also been considered as a useful factor for the speech classification and silence removal. The unique quality of pitch is that it rises when the speech is voiced and zero for silence speech (Sharma and Rajpoot 2013).

There are a number of techniques in the literature that uses time and frequency domain features (e.g. Energy, zero cross rate, spectral centroid etc.) and pattern recognition methods to remove the silence part of speech utterance (Sahoo and Patra 2014), (Zhang 2014), (Liscombe and Asif 2009), (Saha et al. 2005), (Giannakopoulos 2014). Figure 2-4 represents the sequential steps involved for the speech classification into voiced, unvoiced, and silence segments proposed by (Sharma and Rajpoot 2013). The system is tested on a dataset of 15 words spoken by 4 speakers (3 male, 1 female) and achieved an average performance of 97% for speech classification into voiced, unvoiced and silence parts. A test case scenario is presented in Figure 2-5 where the original signal is processed by the aforementioned silence removal approach. It can be analysed that the total number of speech samples are reduced from 85000 to 65000 after removal of silence segments.

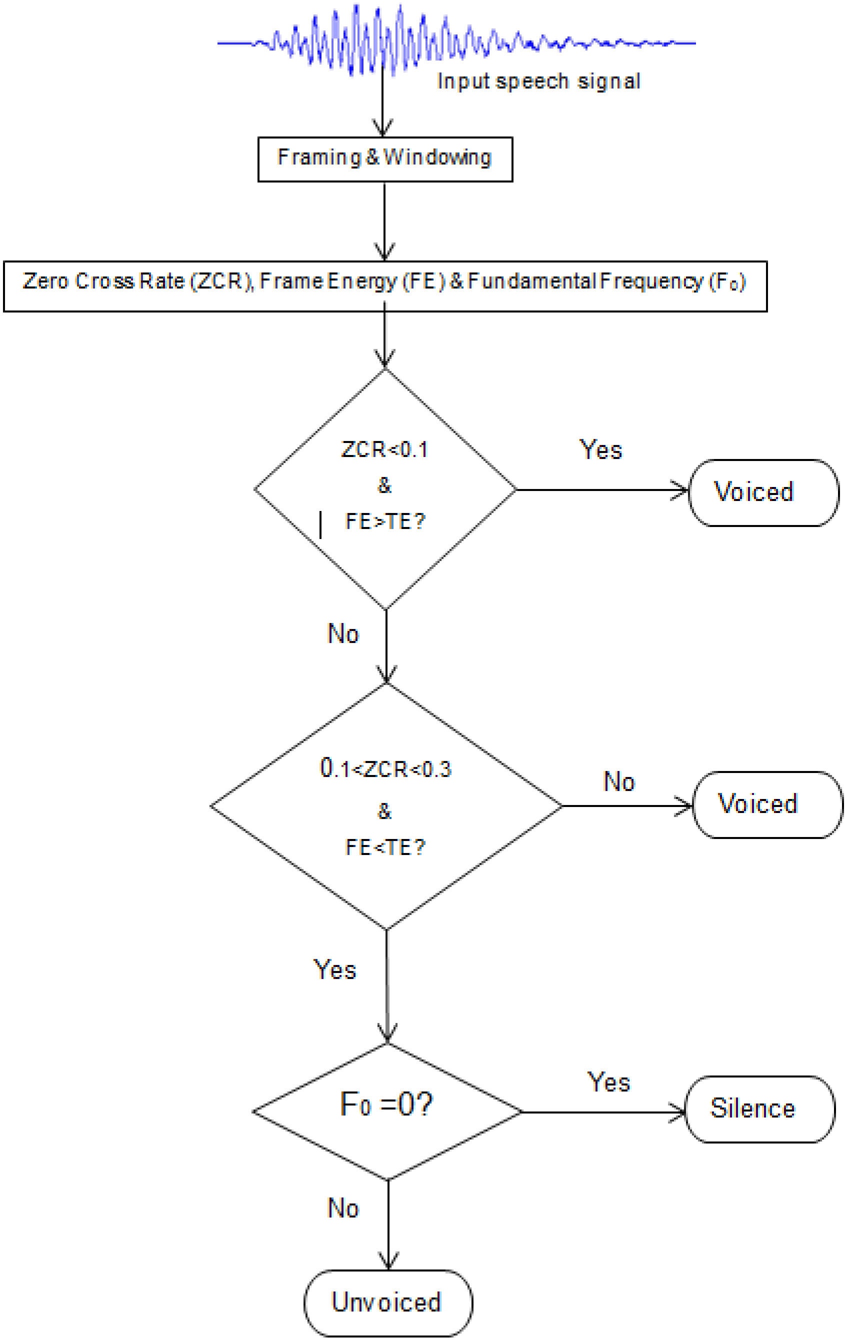


Figure 2-4: Silence Removal Process Using Energy, ZCR and F_0

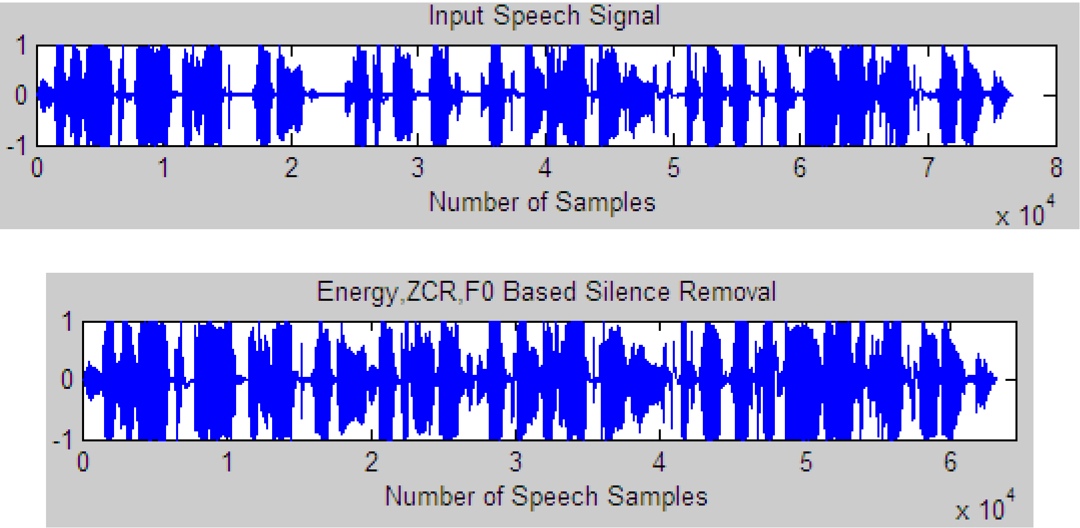
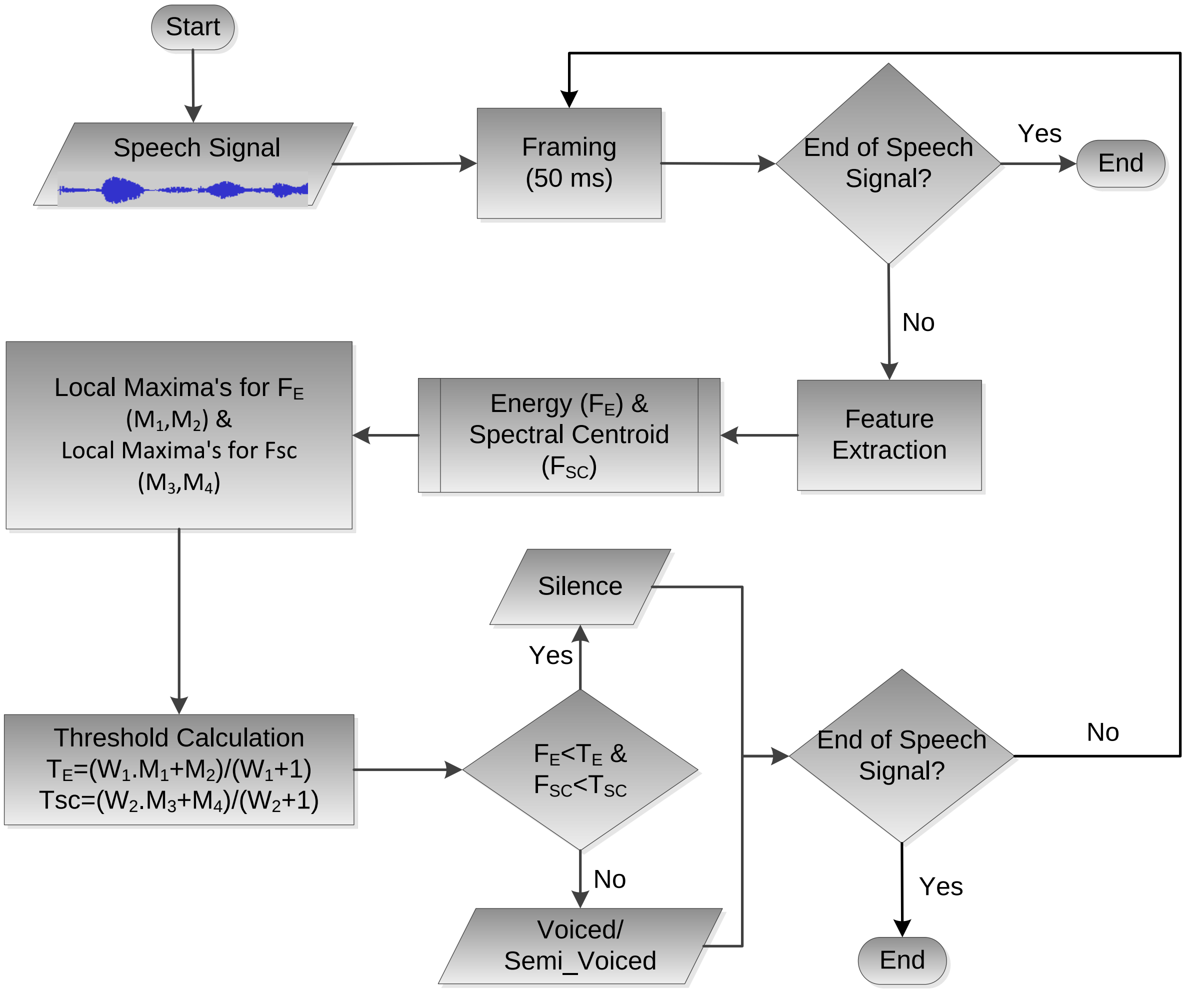


Figure 2-5: Silence Removed By Energy, ZCR and F_0 Based Approach

Similarly, an efficient Voice Activity Detection (VAD) technique is proposed by (Giannakopoulos 2014) based on ZCR and spectral centroid. In order to extract the feature sequences, speech signal is first broken into non-overlapping short-term-windows (frames) of 50 milliseconds duration. Then for each frame, signal energy and spectral centroid is calculated. Along with the two feature calculation, a simple threshold-based algorithm is applied in order to classify the speech segments as shown in Figure 2-6. The process is executed for both feature sequences, leading to two thresholds: T_1 and T_2, based on the energy sequence and the spectral centroid sequence respectively. As long as the two thresholds have been estimated, the threshold value is set for two feature sequence. The segments are formed by successive frames for which the respective feature values are larger than the computed thresholds. Finally, the identified speech segments are extended by 5 short term windows length and merged together to form silence free speech signal.



Start
Speech Signal
Framing (50 ms)
End of Speech Signal?
Yes
End
No
Feature Extraction
Energy (F_E) & Spectral Centroid (F_{SC})
Local Maxima's for F_E (M_1,M_2) & Local Maxima's for Fsc (M_3,M_4)
Threshold Calculation
$T_E=(W_1.M_1+M_2)/(W_1+1)$
$Tsc=(W_2.M_3+M_4)/(W_2+1)$
$F_E<T_E$ & $F_{SC}<T_{SC}$
Yes
Silence
No
Voiced/ Semi_Voiced
End of Speech Signal?
No
Yes
End

Figure 2-6: Time Domain Feature Based Voiced, Unvoiced and Silence Detection

In the proposed research study (Khan and Holton 2015), a novel approach is introduced for the silence removal by deploying a pitch tracking technique introduced in (Zahorian and Hu 2008). Detailed process of pitch tracking and silence removal is presented in the Chapter 4, Section 4.3. Also, keyword spotting and speech tracking performances are compared and presented in performance evaluation sections (Chapter 4, 5, and 6) for the aforementioned silence removal approaches.

2.2.2. Time Frequency Representation and Spectral Analysis

For the analysis and processing purpose, usually speech signal is divided into smaller units called *segments* or *frames*. Segmentation is the very basic step in any voiced activated systems that include speech recognition systems and speech synthesis systems. Speech signal can be segmented into a set of fundamental acoustic units that include words, phonemes or syllables. Most of the time, phonetic units are used for the purpose of speech recognition and synthesis. Because of the slow varying nature of the speech signal, the processing of speech is conducted in blocks or frames of specific length over which the properties of the speech waveform can be assumed as stationary signal.

Extensive research work has been conducted related to frame size selection. It is analysed that speech signal remains almost stationary for the duration of 10 milliseconds to 50 milliseconds (Ravindran et al. 2010) depending upon speaker, speech contents, and nature of the problem. The frame by frame analysis is called short term, time domain or frequency domain analysis of speech signal. Time domain analysis of speech signal provides some dominant features that are useful for different purpose. For example, zero crossing, time domain signal energy, and autocorrelation are some dominant features that may help for a variety of applications. Speech signal contains very rich information that if not lost, may help in diverse application areas. Speech synthesis, speaker recognition, speech recognition, and speech compression are the most common examples of such applications. There are some dominant features existing in speech signal that can’t be extracted in time domain, hence it is necessary to transform the speech signal into frequency domain to retrieve its spectral components. A number of techniques have been used in literature for the purpose of time-frequency representation of speech signal. A Discrete Fourier Transform (DFT), STFT, and WT are the most commonly used spectral analysis techniques (Mayer 1989) that are described in the following sections.

2.2.2.1. Fourier Transformation

Time series data and hence speech signal can be described by a waveform as a function of time. The FT basically decomposes the speech signal into sinusoids to represent in alternative way. The FT provides information about the time varying nature of spectral features existing in the speech signal. The continuous FT converts the signal from time to frequency domain for infinite duration of continuous spectrum that is made up from infinite number of sinusoids. In the context of speech signal, most of the related research work is based upon DFT that is presented below.

A. Discrete Fourier Transform

Suppose $x(t)$ is continuous time signal, then continuous FT is given by:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \qquad 2\text{-}2$$

Where; 'f' is the current frequency being analysed, 't' is a current sample in'x', $e^{-j2\pi ft}$ is the alternative representation of complex trigonometric function that indicates the circular path covered with specific frequency. In case of speech signal processing, we deal with digital signal that is sampled at specific frequency. Hence $x(t)$ is sampled as $x_0, x_1, ...x_{N-1}$ over time '*T*' with Δt as sampling interval. Moreover, continuous FT deals with infinite time '*T*' and number of samples '*N*' which is impossible. To deal with the infinity, the continuous FT can be replaced by DFT. Mathematically:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \qquad k = 1, 2, ..., N \qquad 2\text{-}3$$

With spectrum frequency of $\omega_k = (2\pi)\frac{k}{T}$, X_k represents the DFT of the sampled signal x_n.

B. Short Time Fourier Transform

The major limitation of continuous and DFT is the inability of simultaneous representation of time and frequency localisation. Consequently, DFT is considered not useful for the time variant and non-stationary signal analysis. One possible solution for time-frequency localisation was presented in the form of where input signal is broken into equal length short term frames which normally overlap each other to reduce the artefacts at the boundary. Each frame is windowed that means the multiplication with a smooth function that gradually decreases towards both ends of the frame. A DFT is computed for each overlapped frame and the complex output is added into a matrix which keeps the record of magnitude and phase for each point in time and frequency (Allen and Rabiner 1977). Mathematically, it can be represented as follows:

$$X_m(w) = \sum_{t=-\infty}^{\infty} x(t)w(t-mR)e^{-jwt} \qquad \text{2-4}$$

$$= DTFT_w(x.SHIFTmR(w))$$

$x(t)$ = input signal at time t

$w(t)$ = Hamming window function

$Xm(w)$ = DTFT of windowed data centered about time mR

R = hop size in samples, between successive DTFTs

Where; 'm' is the index for window shifting. Output of STFT is 2-dimensional matrix with '*N'* rows and '*M'* columns that can be represented as:

$$N = WindowSize/2+1, \quad \text{if input signal is real}$$
$$= WindowSize/2, \quad \text{if input signal is not real}$$

$$M = floor\left(\frac{DataSize - WindowSize}{WindowSize - OverlapSize}\right)+1$$

The FT has been used in a wide range of application areas that include differential equations analysis, spectroscopy, quantum mechanics, and signal processing (Allen and Rabiner 1977). In the context of speech, DFT and STFT have been mostly used for time frequency analysis of signal (Dhingra et al. 2013), (Dave 2013), (Lakshmikanth et al. 2014) that is needed for feature extraction and noise reduction in speech signals. As an example; a speaker recognition system is presented by (Farah and Shamim 2013) that use STFT to extract the MFCC feature vectors from the speech signal. Similarly, a key word spotter is presented in (Bahi and Benati 2009) based on STFT as a spectral analysis method. The frequency domain data is processed for the MFCCs calculation while energy of the speech signal is calculated from the time domain speech signal. Both; energy and MFFCs coefficients were used as a feature set that were incorporated for the system training. The system is analysed using hamming window with 36 milliseconds frame length. The VQ algorithm is used for the vector training and codebook generation. Each codebook represents the acoustic features of the signal. HMMs were used for the probability assignment for observation given a word. If the highest probability is up to a given threshold, the system confirms the detection of the keyword. The system is tested on the 10 files each having 10 seconds length. Results show that the system outperforms for a limited dataset with very minor error rate. However, this system is based on VQ and HMM that need a high computational cost. Moreover, HMM based recognizers depend upon probability assignment that may not be assigned correctly and may cause misrecognition.

A sampled speech signal analysis model based on STFT is presented by (Griffin and Lim 1984). Speech signal was assumed as the output of a linear non-stationary function operated by a quasi-periodic impulse. A narrow-band and wideband analyses of voiced speech is obtained using STFT. For the case of wide-band analysis of voiced speech, additional assumption of a formant structure for the speech yielded a simplified interpretation of the wide band STFT. The transformed signal is represented by the features including magnitude, phase, and instantaneous frequency of the STFT. This model is applicable to a wide range of areas specifically, for speech spectrogram that is widely used for speech analysis, speech synthesis, phase Vocoder, channel Vocoder, sub-band Vocoder, and transform Vocoder. Similarly, it has a number of applications for time compression and expansion of speech signal (Portnoff 1978) and (Portnoff 1981).

Research conducted in our previous work (Khan et al. 2012) provides the solution for word similarity matching based on Posterior Probability Measure (PPM). The STFT is applied for the frequency domain transformation of speech signal. The transformed data is passed to PPM frame by frame and a similarity score is calculated. System performance showed that the PPM provides an efficient signal matching performance in the presence of background noise in the speech signal as compared to existing similarity measure algorithms. This approach uses STFT based frequency components for the dominant feature extraction that are processed further for a similarity measurement.

The issue associated with STFT is the limitation of a simultaneous time frequency representation. It can provide a simultaneous time frequency localisation by a compromise on window size that causes the limitation for localisation accuracy. Large size window provides good frequency resolution but poor time resolution and vice versa. This problem is known as *Heisenberg uncertainty principle* which says that the exact time-frequency representation of a signal can’t be known. Consequently, the unnecessary time and frequency components that may cause a mismatch and misidentification cannot be filter out. This issue is resolved in the proposed research work by the usage of WD for time-frequency representation and noise filtration from the speech signal (Khan et al. 2014).

2.2.2.2. Wavelet Transform

In the STFT, sinusoids are used as basic functions and windows are used to localise the signal to a particular time interval. In wavelet transforms, the windows are incorporated directly into the basic functions and a variety of non-sinusoidal shapes are common. “A wavelet (referred as *mother-wavelet*) is a short term signal with an average value of zero”. As compared to sinusoids which works better for stationary signal, Morlet introduced the idea of wavelets (Mayers 1989) that are irregular and works better for non-stationary and non-periodic signals. The wavelet transform uses the mother-wavelet likewise STFT uses a windowed sinusoid. In the wavelet transformation, mother-wavelet is stretched and compressed with the shifting process along the input signal. This process is called scaling and is analogues to frequency in STFT. The mother wavelet is scaled and shifted along the whole signal to calculate the frequency spectra for corresponding scale as shown in Figure 2-7.

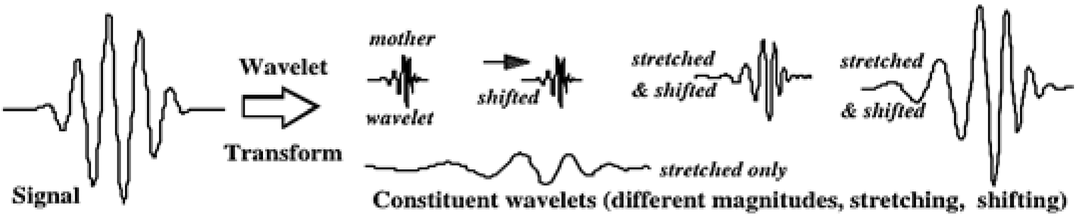


Figure 2-7: Scaling and Shifting Process of Mother Wavelet (Fugal 2009)

Above figure shows the shifting and scaling process of mother wavelet. At first step, a mother-wavelet is picked up and shifted along the whole signal to check the frequency at that level. In the next step, mother-wavelet is stretched and again shifted along the whole signal to get the frequency available at next level. The process of scaling and shifting runs recursively and a three-dimensional time-frequency representation is achieved as shown in Figure 2-8 (c).

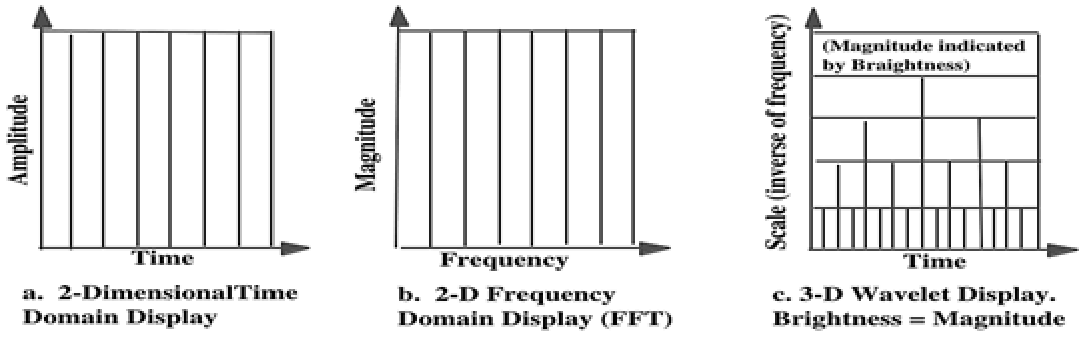

Amplitude
Time
a. 2-DimensionalTime Domain Display
Magnitude
Frequency
b. 2-D Frequency Domain Display (FFT)
Scale (inverse of frequency)
(Magnitude indicated by Braightness)
Time
c. 3-D Wavelet Display. Brightness = Magnitude

Figure 2-8: Difference between FFT and Wavelet Transform (Fugal 2009)

Above figure describes the time-frequency representation difference between STFT and wavelet transform. The STFT introduced short time window concept that solved the time localisation issue to some extent with a compromise on window length. Wavelets, on the other hand solved the both issues by introducing the mother-wavelet shifting and stretching over time, to provide time-frequency domain analysis while magnitude of the frequency as 3rd dimension as shown in the Figure 2-8 (c). Mathematically;

$$CWT_x^{\Psi}(\tau,s)=\Psi_x^{\Psi}(\tau,s)=1/\sqrt{s}\int x(t)\Psi^*(t-\tau/s)dt \qquad 2\text{-}5$$

The above equation provides the frequency transformation of the speech signal using continuous wavelet transform as a function of *shift* and *scale* parameters represented by 'τ' and '*s*' respectively. The transforming mother-wavelet is represented by $\Psi(t)$, 't' is the time axis and output is known as Continuous Wavelet Transform (CWT). The 'τ' corresponds to the time information of the window whereas, '*s*' indicates different level of frequencies in the signal. Larger value for '*s*' corresponds to dilated signal while smaller value corresponds to compressed signal.

The transformation process of speech signal using wavelets is represented in the figure below.

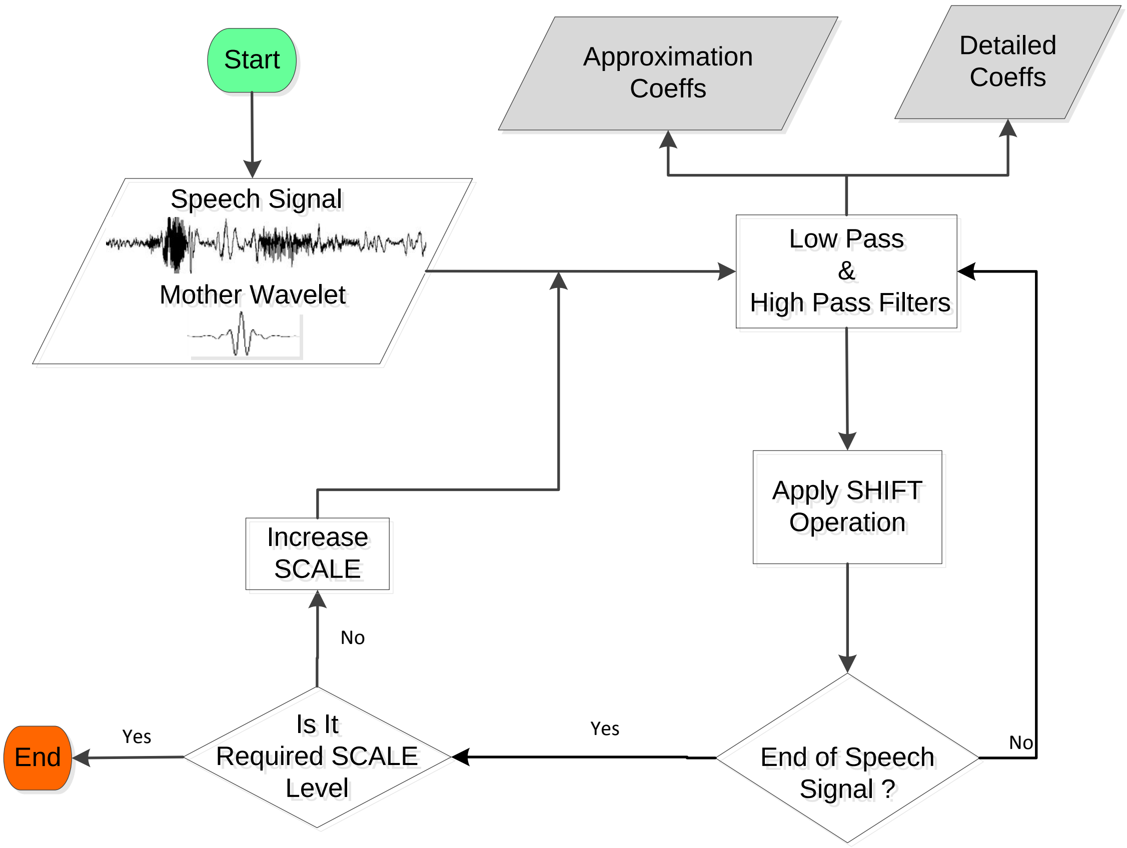


Figure 2-9: Wavelet Transform of a Speech Signal

Discrete Wavelet Transform (DWT) on the other hand uses a dyadic set of scales that means; the input signal is decomposed into mutually orthogonal set of wavelets which is the main difference from a CWT. The DWT is based on smoothing and non-smoothing filters that are constructed from wavelet coefficients. Suppose '*L*' *is* the signal length with 2^N number of sampled data '*D*' then at first step; *D*/2 data at scale $L/2^{(N-1)}$ are computed. Then (*D*/2)/2 data at scale $L/2^{(N-2)}$ and so on till last 2 data at scale *L*/2 as shown in figure below.

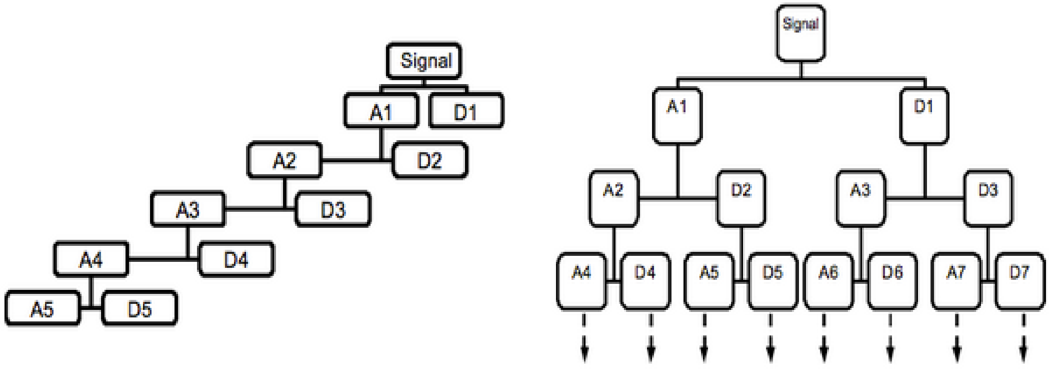


Figure 2-10: WD into Approximation and Detailed Coefficients (Meyer 1989)

In Figure 2-10, $A_{1,2,3\ldots}$ and $D_{1,2,3\ldots}$ represents the approximation and detailed coefficients at level 1,2,… respectively. The input signal can be reconstructed by adding up the approximation and detailed coefficients at each level. The major difference between WD and wavelet packet decomposition is shown in the Figure 2.10. It is clear that wavelet packet decomposition (right side figure) expands both; approximation and detailed coefficients into '*A*' and '*D*' whereas WD does expand only approximation coefficients in the next level. There are different families of mother-wavelet with corresponding properties in terms of structure as shown in Figure 2-11.

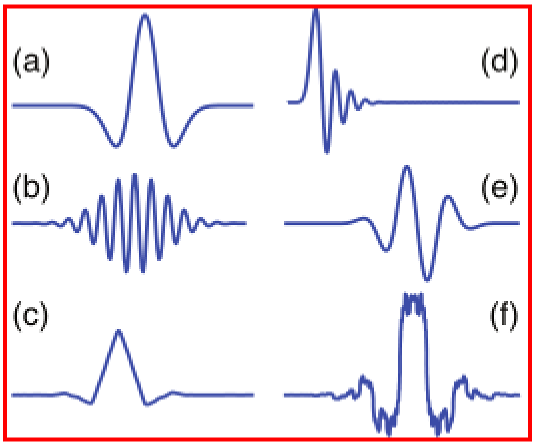


Figure 2-11: (a) Mexican Hat Wavelet, (b) Complex Morlet Wavelet, (c) Coiflets Wavelet, (d) Daubechies Wavelet, (e) Complex Gaussian Wavelet, and the (f) Bi-Orthogonal Spline Wavelet (Fugal 2010).

Wavelet transform have continuously and successfully been implemented in the literature for diverse application areas. For example, wavelets have an extensive use for the data compression, time varying signal analysis, noise filtration, and signal smoothing, speech recognition and word identification applications (Akansu et al. 2010). In addition, wavelet transform has been used in the area of communication. Orthogonal Frequency Division Multiplexing (OFDM) is one of the best examples of such applications. Wavelet based OFDM is capable to perform efficiently in terms of deep notches achievement as compared to FFT based OFDM (Galli and Logvinov 2008).

Over the past two decades, research works have been directed towards the use of wavelet based feature extraction (Byung-chul 2001), (Tufecki 2001), (Sarikaya 2001) in the area of ASR and similarity measure. Three dimensional representation property of DWT can be used for the filtration for the signal of interest. For example, a wavelet based speech recognition is presented by (Gamulkiewicz and Weeks 2003) for the phonemes matching. A DWT (Daubechies 8) is applied on the pre-processed data to obtain five levels of frequency bands. User speaks a test word for the recognition. Using DWT, the test word is converted into five frequency bands that are compared with the template phonemes frequencies to generate the identified phonemes list. The system is tested on a small dataset of 35 template phonemes with 5 recordings for each phoneme as test data. The performance results showed overall 57% correction out of total 175 possible matches. Moreover, it is observed that the best performance (77%) is achieved using first three approximation coefficients as compared to higher frequency bands.

A speaker identification system is presented by (Shafik et al. 2009) on the basis of wavelet transform based MFCC feature vectors that works efficiently in the presence of different level of noise in telephonic speech signal. The noisy speech signal is passed to DWT that transforms the signal from time to frequency domain by breaking it up into approximation and detail coefficients. Both; approximation and detailed coefficients are concatenated and passed to MFFC features extraction process. The MFCC coefficients of the original noisy signal were extracted too. Both MFCCs feature vectors were concatenated to produce a huge feature space for the speaker identification. The system is trained using ANN over the extracted features from a dataset of 150 Arabic sentences by 15 speakers. For the test purpose, same speakers were asked to repeat a sentence with the same contents. Different scenarios have been generated for the system performance evaluation based on time domain MFCCs, DWT based MFCCs, and combination of both. It is analysed that the best performance is achieved when the system is trained on the combined MFCCs feature vector produced by DWT and original signal. Moreover, system performed efficiently not only in low level of SNR but also in moderate and high level of SNR.

An effective research work is presented by (Fugal 2010) that use Dabuchies wavelet for finding the matching patterns in speech waveform in the regional dialects of demographic region. Speech samples were recorded by a number of speakers from five different dialects regions. The data is loaded to wavelet filter for the transformation into frequency domain. Standard deviation, mean, variance and mode were used as the feature set for the similarity calculation between test and template waveform. Although, the contents of speech test and template speech were different, yet the performance result showed a huge resemblance in the corresponding dialect speech contents. In the proposed research work, WD is used as a spectral analysis and filtration method for the continuous speech tracking and keyword spotting applications. A detailed analysis of the WD based dynamic noise filtration and feature extraction performance is presented in Chapter 5.

2.2.3. Feature Extraction

Feature extraction is a common term used in pattern recognition area that means the measurement for characteristics (i.e. features) of a specific pattern (e.g. speech signal in proposed research). These features are passed to a classifier or similarity measure algorithm to identify the pattern (e.g. speech utterance). Basically, a speech signal consists of bunch of information that is useful for multiple purposes including speech recognition, word identification, and speaker recognition etc. This information exists in the form of variety of acoustic features in speech signal that represent the dominant properties of a specific segment of speech signal. Moreover, as discussed in previous section, speech signal can be considered as a stationary in short term analysis. These features are considered constant for that segment of short time (10-50 milliseconds) (Ravindran et al. 2010). Features may be extracted in time domain as well as in frequency domain. In the literature, Mel Frequency Cepstral Coefficient (MFCC) is the most dominant feature set that have been used for the automated speech recognition (Dhingra et al. 2013), (Dave 2013). However, there are some other features that may be useful for other purposes related to speech processing that include Linear Predictive Coding (LPC) (Bradbury 2000, Juricka 2014), Perceptual Linear Predictive (PLP) (Dave 2013), (Hermansky 1990), energy, ZCR, and F_0 (Zahorian and Hu 2008). In the proposed research work, MFCC and WD based Energy are used as feature set for the purpose of TWCST and keyword spotting.

Mel Frequency Cepstral Coefficient

In speech signal processing, elimination of the unnecessary components from speech signal is one of the challenging task. These components may be in the form of background noise, silence, and emotion etc., that if not eliminated, may lead to misidentification or misrecognition. As discussed earlier, sound generated by human are filtered by vocal tract and other linguistic parts including jaws, nose, and tongue etc. On the basis of this filtration, shape of the output sound is determined in the form of phonemes. The vocal tract shape produces the output sound in the form of short time power spectrum envelopes that are also produced by MFCC. The MFCCs were first time introduced by Davis and Mermelstein in the 1980's and have been used in ASR and speaker recognition area as state-of-the-art ever since (Lyons 2012), (Farah and Shamim 2013).

Basically, vibrations of the human cochlea at different spots depending on the frequency of incoming sounds in the form of power spectrum, cause different nerves fire to produce information to brain about the existence of certain frequencies. The spectrogram of speech signal performs same task of identifying which frequencies are present in the frame. As there are unnecessary frequency components in the spectrogram of signal, it is hard to classify two closely related frequencies. For this purpose, the whole spectrum is divided into different frequency regions and the corresponding energy is calculated for individual ranges. In the context of MFCC, this is performed by Mel filter-bank (Lyons 2012). The Mel filter-bank filters are gradually stretched from start to end to indicate the energy levels for zero towards higher frequencies respectively. Hence, the Mel scale provides a range measure for the width of a filter-bank. In the next step, the filterbank energies are logged to make it more closely to what human hears. This is because human hearing doesn’t work on linear scale. Finally, Discrete Cosine Transform (DCT) of log filterbank energies is computed to de-correlate the energies.

To process the input speech for MFCC feature extraction, firstly, segmentation is applied to input speech signal with frame length of 20-50 milliseconds. Hence, for a typical sampling frequency of 8 KHz, one frame consists of 0.025*8000 = 200 samples. Then, the DFT is applied on each overlapped frame which is given by:

$$S(k) = \sum_{n=1}^{N} s(n)h(n)e^{-j2\pi kn/N} \qquad 2\text{-}6$$

Where; $h(n)$ is '*N*' sample long analysis window (e.g. Hamming window), 'k' is the current frequency to be analysed by DFT, and 'n' is the sample index. The period-gram based power spectral estimate for the speech frame $s_i(n)$ is given by:

$$P_i(k) = \frac{1}{N} | S_i(k) |^2 \qquad 2\text{-}7$$

The above relation represents the period-gram estimate of the power spectrum. After the power spectrum estimate, Mel-spaced filter-bank is computed that consists of a set of 20-40 (26 is standard) triangular filters that are applied to the period-gram power spectral. These filter-banks comes in the form of 26 vectors of length 257 (depending upon DFT setting for FFT points). Normally, all vectors consists zero values mean with a non-zero for a certain section of the spectrum. Each filter-bank is multiplied with the power spectrum and the coefficients are added up to provide filter-bank energies. As a result, there are only 26 numbers each indicating energy level in the corresponding filterbank. To get log filterbank energies, a log of each of the 26 filterbank energies is taken. In the next step, the DCT is applied to the 26 log filter-bank energies to get 26 cepstral coefficients. In the literature, only lower 12-13 of the

26 coefficients are kept for ASR and speaker recognition application (Lyons 2012). The resulting features (12 numbers for each frame) are called Mel frequency cepstral coefficients.

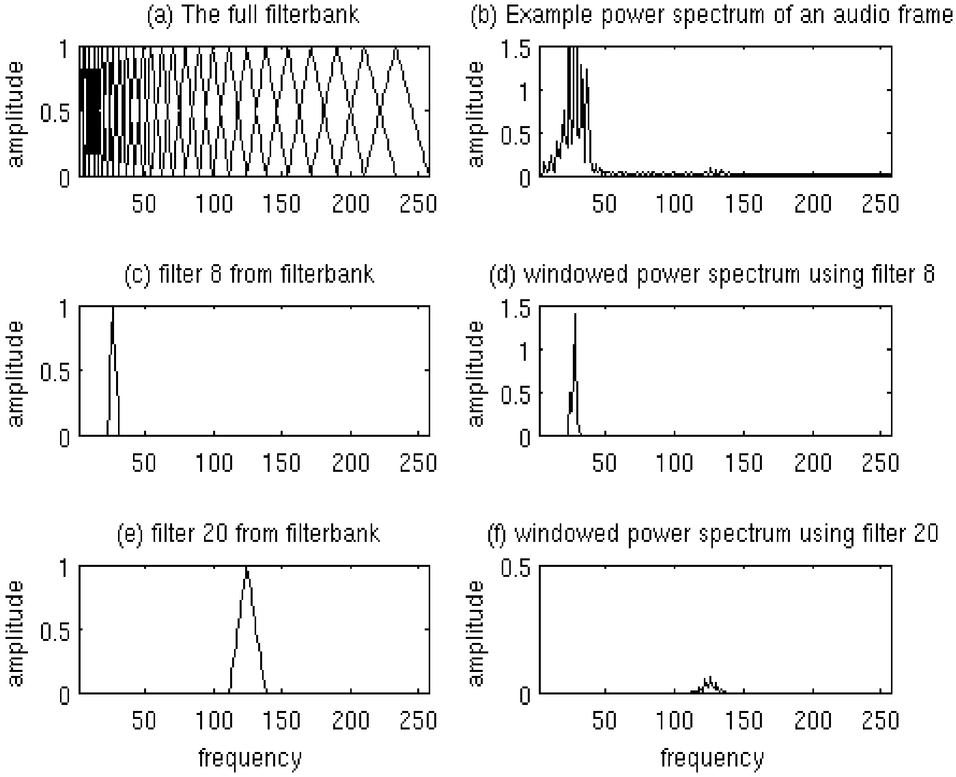


Figure 2-12: Plot of Mel Filterbank and Windowed Power Spectrum (Lyons 2012)

Mel Filter-bank Computation

To calculate the filter-banks shown in Figure 2-12(a), lower and upper frequencies are selected. In most of the literature, lower frequency limit is set as 300Hz and 8000Hz for the upper frequency (Lyons 2012). Moreover, 8 KHz sampling frequency means 4 KHz limit for upper frequency using Nyquist sampling theorem. The upper and lower frequency can be converted into Mels using:

$$M(f) = 1125\ln(1 + f/700) \qquad 2\text{-}8$$

Thus using the above relationship between frequency ‘ f ’ and mel scale, 300 Hz is represented by 401.25 mel and 8000Hz are 2834.99 mel . In case of 10 filterbanks, 12 points are needed that can be added linearly between 401.25 and 2834.99. To convert mel back to frequencies, we use:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \qquad \text{2-9}$$

In the next step, as the frequency resolution required putting filters at the exact points which are not specified, the above frequencies are rounded to the nearest FFT bin depending upon FFT size and sampling frequency. Finally, the filter-banks are created starting from the first point; reach its peak at the second point, then return to zero at the 3rd point. The second filter-bank will start at the 2nd point, reach its max at the 3rd, then be zero at the 4th etc., as shown in Figure 2-13. Equation 2-10 presents a relationship to calculate these boundaries as:

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)} & k < f(m-1) \\ & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ & k > f(m+1) \end{cases} \qquad \text{2-10}$$

Where 'm' is the number of desired filters, and 'f' is the list of '$m+2$' mel spaced frequencies.

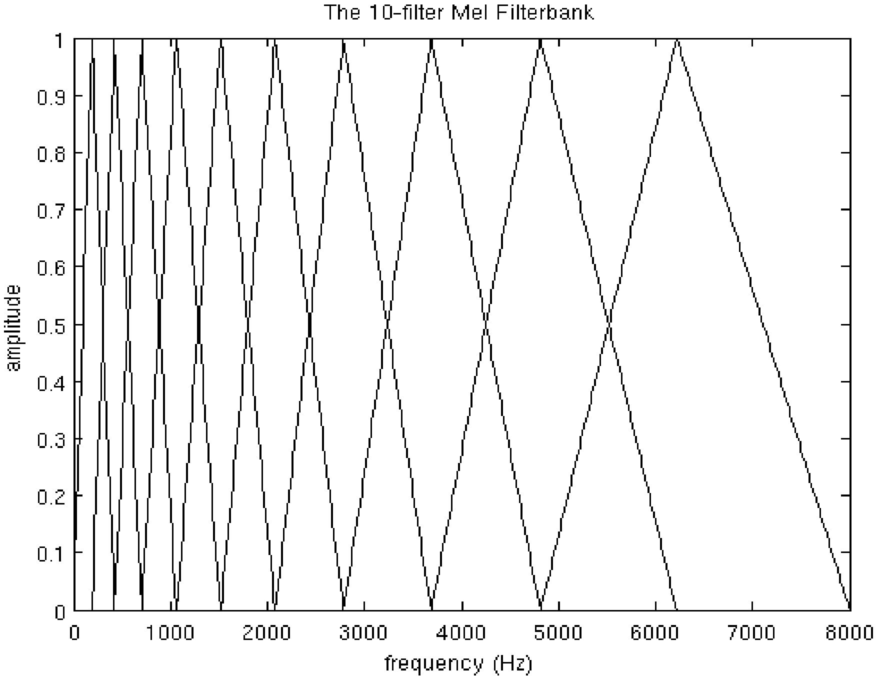


Figure 2-13: A Mel-Filter-Bank Containing 10 Filters (0Hz to 8000Hz), (Lyons 2012).

The MFCC has been used in a variety of applications in the literature. As an example; a key word spotter is presented by (Bahi and Benati 2009) based on MFCC as features vector. The frequency domain data is passed for the MFCCs calculation while energy of the speech signal is calculated from the time domain speech signal. Both; energy and MFFCs coefficients were used as a feature set for the system training. The system is analysed using Hamming window with 36 milliseconds frame length. The VQ algorithm is used for the vector training and codebook generation. Each codebook represents the acoustic features of the signal. The HMMs were used for the probability assignment for observation given a word. If the highest probability is up to a given threshold, the system confirms the detection of the keyword. The system is tested on the 10 files each having 10s length. The results show that the system outperforms for this limited dataset with very minor false detection rate. However, this system is based on VQ and HMM that have high computational cost due to training process. Moreover, HMM based recognizers depend upon probability assignment that may not be assigned correctly and may cause misrecognition.

A speaker identification system is presented by (Shafik et al. 2009) on the basis of wavelet transform based MFCC feature vectors that works efficiently in the presence of different level of noise in telephonic speech signal. The noisy speech signal is passed to DWT that transforms the signal from time to frequency domain by breaking it up into approximation and detail coefficients. Both; approximation and detailed coefficients are concatenated and passed to MFFC features extraction process. The MFCC coefficients of the original noisy signal were extracted too. Both MFCCs were concatenated to produce a huge feature space for the speaker identification. The system is trained using ANN over the extracted features from a dataset of 150 Arabic sentences by 15 speakers. For the test purpose, same speakers were asked to repeat a sentence with the same contents. Different scenarios have been generated for the system performance evaluation based on, time domain MFCCs, DWT based MFCCs and combination of both. It is analysed that the best performance is achieved when the system is trained on the combined MFCCs feature vector produced by DWT and original signal. Moreover, system performed efficiently not only in low level of SNR but also in moderate and high level of SNR.

Similarly, (Farah and Shamim 2013), (Dhinjra 2013), (Gandhiraj and Sathidevi 2007) and (Katsamanis 2009) have used MFCC for speaker identification, speech recognition, and speech emotion recognition resulting a considerable performance. Meinard (2007) utilized MFCC for gender classification and audio similarity measure. Along with the advantages of MFCC, there is a limitation associated in the form of their use in the presence of additive noise. Researchers have been working to overcome this limitation and successfully improved the MFCC performance by increasing the log-Mel amplitude power (Tyagi and Wellekens 2005). In the proposed research, MFCCs coefficients are extracted and used in different forms as the dominant features for the TWCST and keyword spotting approaches.

2.3. Summary

In this chapter, a detailed study on the concepts of speech signal processing is presented. The entire speech signal processing is broken down broadly into different components. This includes speech enhancement and pre-processing, speech segmentation, spectral analysis, and feature extraction. Each component is further explained in depth along with the formulations and applications in the literature and in the proposed research work. In the pre-processing and speech enhancement component, more emphasize was on most commonly used techniques for the silence removal. Segmentation section provides a deep review on framing analysis. Spectral analysis (i.e. time frequency representation) is the most important section of this chapter that provides a complete sketch of time-frequency representation methods in terms of literature review, implementation methodologies, and exploitation in the proposed research work. More specifically, wavelet transforms and Fourier transform are discussed in detail. The formulation of MFCC based features vector extraction from speech signal is addressed in the final section.

3. RELATED WORK FOR SPEECH SIMILARITY MEASURE AND KEYWORD SPOTTING

3.1. Scope

This chapter consists of four sections which aim to provide a comprehensive insight into: (1) dynamic time warping, (2) Kalman filter, and (3) the current level of speech processing related research performed using these approaches. The first section introduces similarity metrics followed by a detailed description of DTW including its implementation and constraints in Section 3. Section 4 demonstrates the most recent research work that is introduced in the area of keyword spotting, isolated word recognition based on DTW, spoken term detection, and continuous speech matching. In the final section, a Kalman filter is introduced and its related work in the area of speech processing is discussed.

3.2. Introduction

Similarity measure is a function that provides the quantity of similarity between two objects. In terms of speech signal matching, it provides a measure of how much two utterances are similar to each other. Another most commonly used terminology that is the inverse of similarity measure is known as distance metric. Distance metric provides the quantity of dis-similarity between two objects. There are more than one definition for similarity measures and distance metrics with respect to their area of applications. Literature consists of a vast use of distance metrics in the area of pattern matching and time series data similarity measurement. The most widely used distance metrics include Euclidean distance (Carlin et al, 2011), Vector Cosine Angle Distance (VCAD) (Yuan and Sun 2005), (Hafner et al. 1995); Bhattacharyya coefficient (Sahoo and Patra 2014), (Djoudi 1990), (Lin 1991); Kullback-Leibler divergence (Liu and Chen 2004); normalized cross correlation (Theodoridis and Koutroumbas 2003); histogram intersection distance (Joukhadar et al. 1999), Tanimoto coefficient (Willet et al. 1998), and a PPM (Fing 2008). The cross correlation and Bhattacharya coefficient are most common pattern matching techniques applied in speech recognition and image processing (Fing 2008). Unlike PPM, the existing similarity measure algorithms lack the ability to differentiate between speech signal and background noise which participates actively in mismatching or misidentification. This issue was addressed in the current research work (Khan et al. 2012) by applying PPM for isolated keyword matching to degrade the effect of background noise in speech signal.

In signal based word identification systems, cross correlation and Euclidean distance are typically used to determine similarities between a pair of spoken words. In document retrieval systems on the other hand, distance metrics based on cosine angle are more commonly used for determining similarity between two documents (Dehak et al. 2010). Even though the Euclidean distance and the cosine angle based distances coincide when the components of the feature vectors are normalised by the norm of the vector, they differ when they are normalised otherwise. In speech processing applications, components of a feature vector are usually normalized by the size of the spectrogram and as a result, Manhattan distance, Euclidean distance, cosine angle based distance, and histogram intersection distance produce degree of similarities between two signals. As in the proposed research study, main focus is on the time warped speech utterance match, therefore; formulation of the Euclidean distance, DTW, and their applications for speech matching and keyword spotting are presented in the following sections.

3.3. Euclidean Distance

In DTW techniques, a Euclidean distance is commonly used as a similarity measure. In a two dimensional coordinate system, if A = (a_1, a_2) and B = (b_1, b_2) are two points, then Euclidean distance between A and B can be defined as:

$$E_{AB} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \qquad 3\text{-}1$$

In case of *n*-dimensions where A= (a_1, a_2, a_3,……, a_n) and B = (b_1, b_2, b_3,……, b_n), then the generalised form of above equation will be:

$$\begin{aligned} E_{AB} &= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots\ldots + (a_n - b_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \end{aligned} \qquad 3\text{-}2$$

The importance of Euclidean distance is based on its fundamental properties of metric. These properties include non-negativity, symmetry, and inequality (Cai and Ng 2004). Most of the time, a metric function is desired because to prune the index during search, the triangle inequality may be used that allows the execution to be speed-up for exact matching (Mueen 2009).

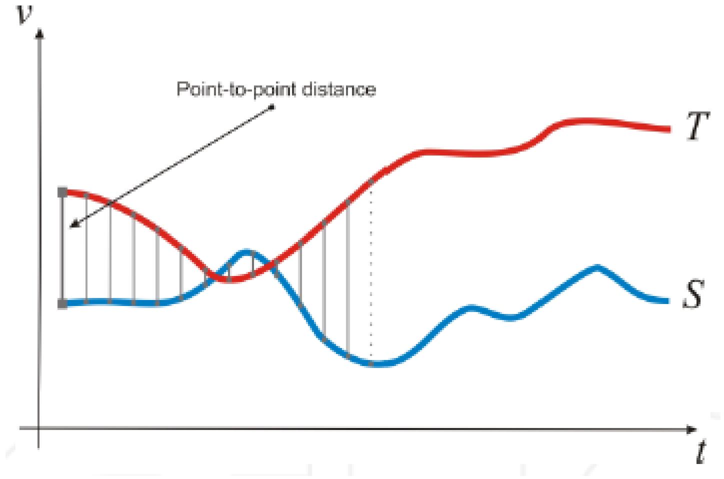


Figure 3-1: T And S Are Two Time Series Of A Particular Variable v Along The Time Axis t.

(Cassisi et al. 2012)

A limited use of Euclidean distance is available related to speech utterance matching. Thakur and Sahayam (2013) proposed an MFCC based speech recognition for password detection based on weighted Euclidean distance. Similarly, feature detection for stress existence in speech signal based on Euclidean distance is proposed by (Ruzanski 2006). The performance is compared with K-means clustering that provided 50.5% error rate as compared to 38% of Euclidean distance based approach. Even Euclidean distance provided better performance, yet the error rate is unreliable in real time applications. The functionality of Euclidean distance measure is affected in a situation where it calculates the in-normalised vectors similarity. As an example, in Figure 3-2, vector ‘A’ and ‘C’ are more similar in terms of direction as compared to ‘A’ and ‘B’, but Euclidean distance results ‘A’ and ‘B’ more similar because of the magnitude difference of vector ‘B’ and ‘C’. The problem can be solved by normalising the vectors which means all vectors will be converted to unit vectors.

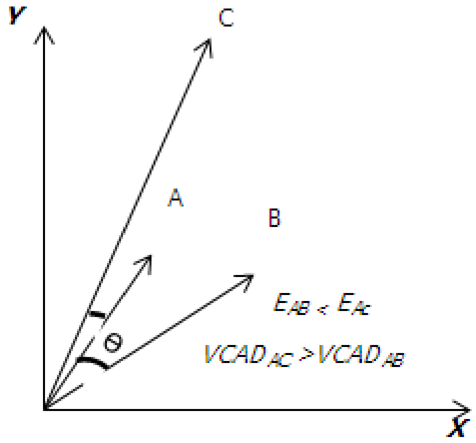


Figure 3-2: Euclidean Distance vs. Cosine Similarity

Another issue related to Euclidean distance is the constraint on query vectors length to be same that does not happen in most of the cases; especially when matching two continuous, non-stationary time series signals. In such a case; DTW is considered one of the best solutions for the time warped signal matching (Carlin et al. 2011), (Ratanamahatana 2002). As compared to Euclidean distance that can deal only with same length sequences, the DTW can measure the distance between two sequences of different length as shown in figure below.

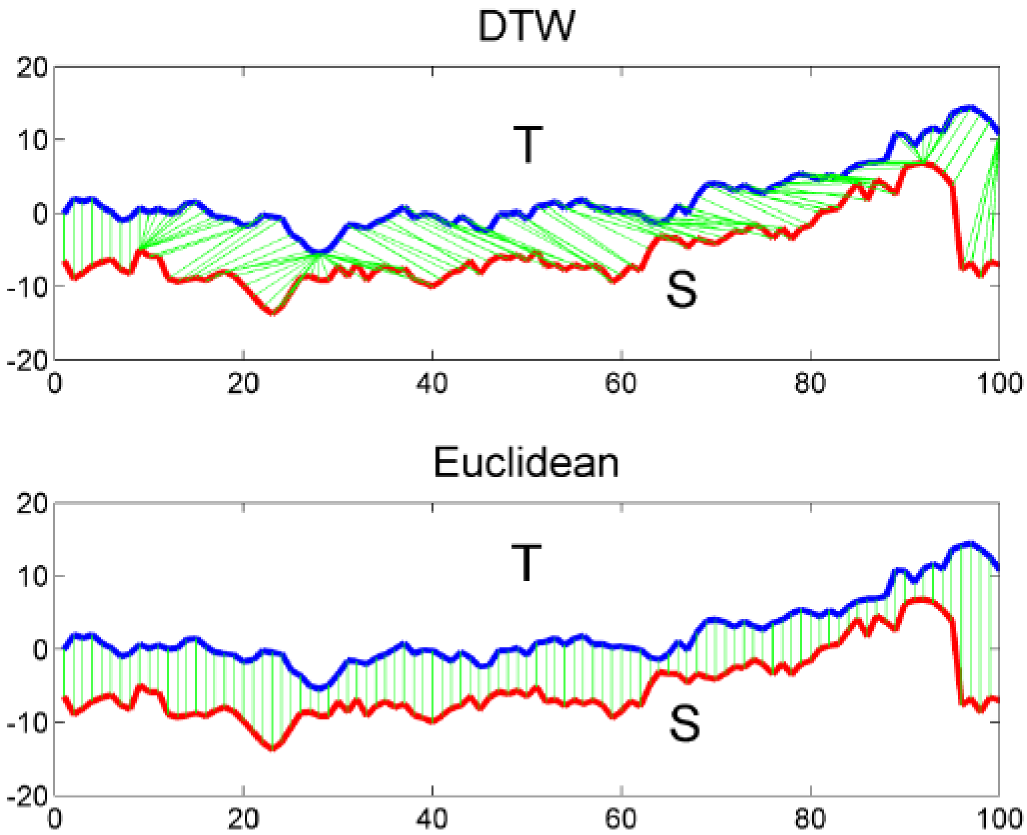


Figure 3-3: Difference between DTW Distance and Euclidean Distance (Cassisi 2012).

3.4. Dynamic Time Warping

The DTW is a well-known technique to find an optimal alignment between two given (time-dependent) sequences that may vary in time. The DTW has been used as one of the powerful algorithm for time series alignment and isolated word recognition (Cassisi 2012). In terms of Spoken Term Detection (STD) and Query by Example (QbyE) related work, the DTW have been used extensively since last decade (Chan and Lee 2010), (Thambiratmann and Sridharan 2007), (Zhang and Glass 2011), (Carlin et al. 2011), (Zhang and Glass 2010), (Jansen et al. 2010), (Zhang et al. 2012). In speech signals, multiple recordings of a speech utterance spoken at different time may differ in length. Moreover, the length of phonemes within the utterance may also differ that causes the time warping issue.

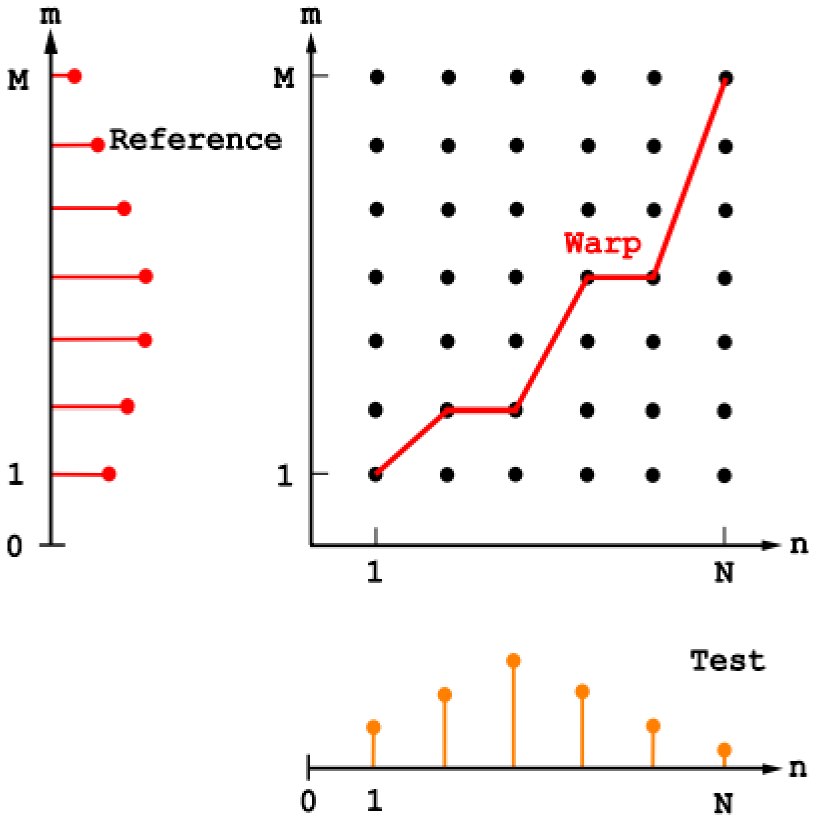
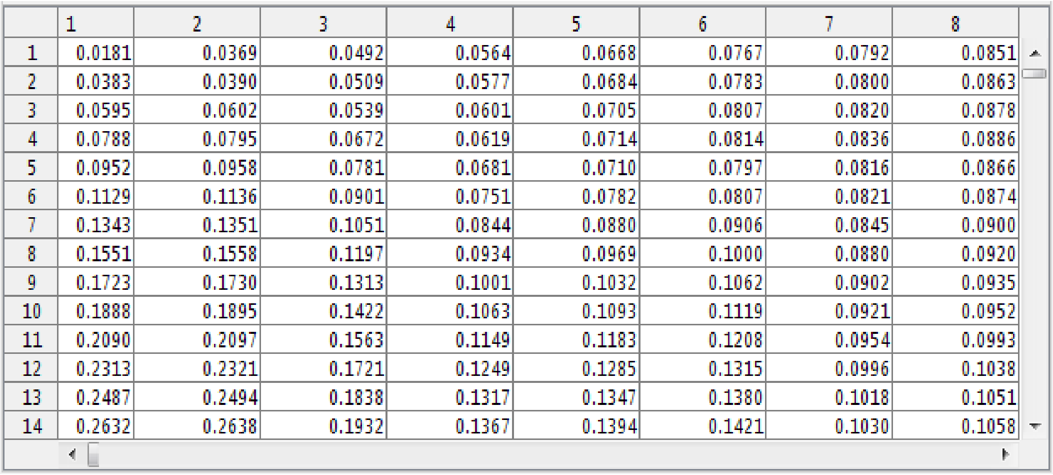


Figure 3-4: DTW Time Axis Alignment (MIT 2003)

In case of two time warped speech utterance matching, the DTW iteratively warps the time axis until an optimal match between two speech utterances is found as shown in Figure 3-4. Usually, the speech signal can be represented as a sequence of its features. The total distance between two sequences is a sum of the minimised local distances. Suppose T = {t_1, t_2, . . . , t_n} and S = {s_1, s_2, . . . , s_m} are two time series of lengths '*n*' and '*m*' respectively, DTW exploits information contained in a '*n* × *m*' distance matrix:

$$distMatrix = \begin{bmatrix} d(T_1,S_1) & d(T_1,S_2),......d(T_1,S_m) \\ d(T_2,S_1) & d(T_2,S_2),......d(T_2,S_m) \\ & . \\ & . \\ d(T_n,S_1) & d(T_n,S_2),......d(T_n,S_m) \end{bmatrix}$$

where $distMatrix(i, j)$ corresponds to the distance $d(T_i, S_j)$ of i^{th} point of ‘T’ and j^{th} point of ‘S’ with 1 ≤ i ≤ *n* and 1 ≤ *j* ≤ *m*. Figure 3-5 provides a sketch of a sub-matrix that provides distances between two time warped isolated words ‘*hello’*.



	1	2	3	4	5	6	7	8
1	0.0181	0.0369	0.0492	0.0564	0.0668	0.0767	0.0792	0.0851
2	0.0383	0.0390	0.0509	0.0577	0.0684	0.0783	0.0800	0.0863
3	0.0595	0.0602	0.0539	0.0601	0.0705	0.0807	0.0820	0.0878
4	0.0788	0.0795	0.0672	0.0619	0.0714	0.0814	0.0836	0.0886
5	0.0952	0.0958	0.0781	0.0681	0.0710	0.0797	0.0816	0.0866
6	0.1129	0.1136	0.0901	0.0751	0.0782	0.0807	0.0821	0.0874
7	0.1343	0.1351	0.1051	0.0844	0.0880	0.0906	0.0845	0.0900
8	0.1551	0.1558	0.1197	0.0934	0.0969	0.1000	0.0880	0.0920
9	0.1723	0.1730	0.1313	0.1001	0.1032	0.1062	0.0902	0.0935
10	0.1888	0.1895	0.1422	0.1063	0.1093	0.1119	0.0921	0.0952
11	0.2090	0.2097	0.1563	0.1149	0.1183	0.1208	0.0954	0.0993
12	0.2313	0.2321	0.1721	0.1249	0.1285	0.1315	0.0996	0.1038
13	0.2487	0.2494	0.1838	0.1317	0.1347	0.1380	0.1018	0.1051
14	0.2632	0.2638	0.1932	0.1367	0.1394	0.1421	0.1030	0.1058

Figure 3-5: Distance Matrix between Times Warped Speech Utterances

The main focus in the DTW is to find the minimum warping path P = {p_1, p_2, . . ., p_k} of contiguous elements on ‘$distMatrix$’ with max(*n, m*) < P < *m* + *n* -1, and w_p = $distMatrix(i, j)$ such that it minimizes the following function:

$$DTW(T,S) = \min\left(\sqrt{\sum_{p=1}^{K} w_p}\right) \qquad 3\text{-}3$$

The warping path is subject to several constraints (Ratanamahatana 2002). Given w_p = (*i, j*) and w_p -1 = (*I’, j’*) with *i, i’* ≤ *n* and *j, j’* ≤ *m*.

Boundary conditions: In DTW, the warping path starts from bottom left and ends at top right; i.e. p_1 = (1,1) and p_K = (*n, m*). Figure 3-4 demonstrates the example of warping path progression that starts from bottom left and progress towards the top right corner of the matrix.

Continuity: Progress in the warping path is made one step (depends upon constraint) at a time (Figure 3-6). Both *i* and *j* can only increase by at most 1 (depends upon constraint) on each step along the path, i.e. *i – i’* ≤ 1 and *j – j’* ≤ 1.

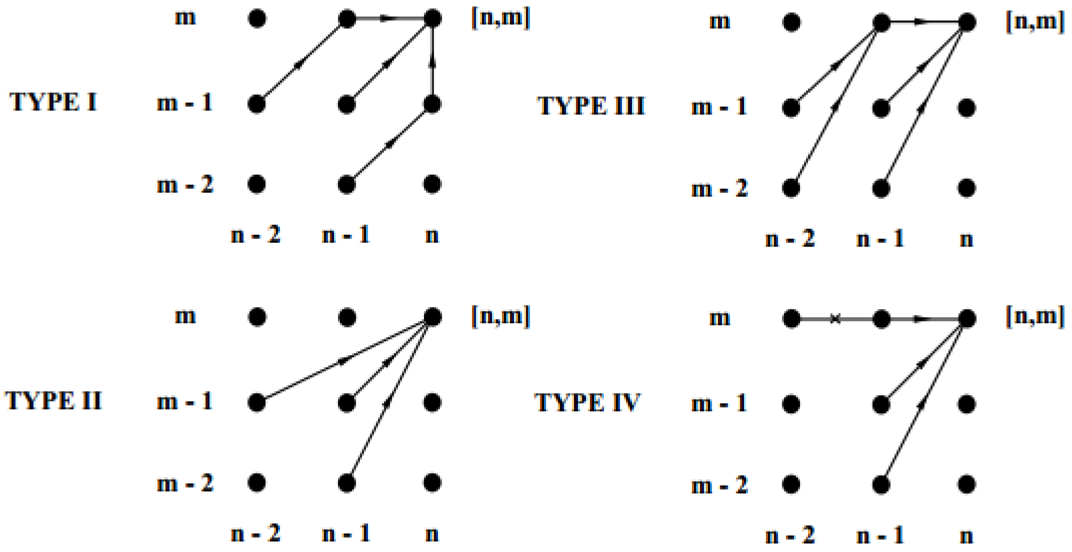


Figure 3-6: Local Constraints and Alignment Flexibility (MIT 2003)

Monotonicity: The path will progress forward only. It can’t move in backward direction even a single step.

Warping width: A path is considered optimal if it doesn’t move very far from the diagonal as shown in Figure 3-7. The distance that the path is allowed to extend is known as *warping width*.

Slope constraint: The optimal path never provides too steep or too shallow slope. A big advantage of this restriction is the prevention of short sequences matching too long ones.

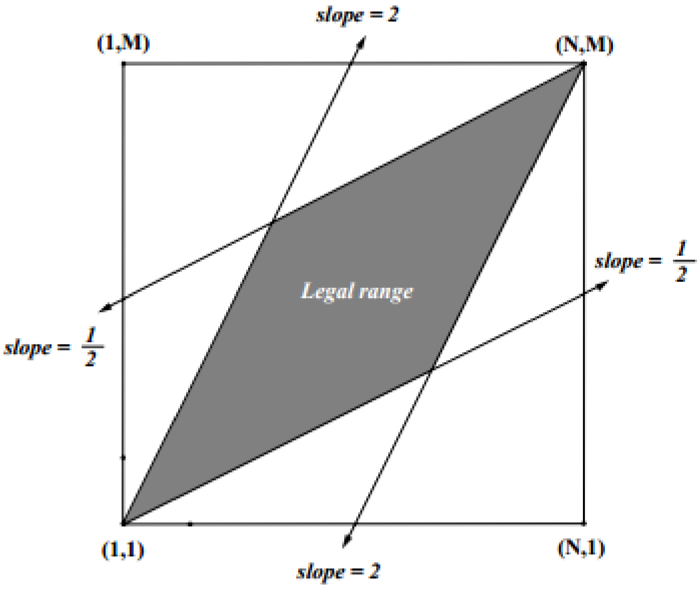


Figure 3-7: Global Constraint: Exclusion of Search Space Using Local Constraints (MIT 2003)

Dynamic programming (DP) is used to calculate the required warping path by generating a cumulative distance metric using;

$$\gamma(i,j) = d(T_i, S_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \qquad 3\text{-}4$$

Where 'γ' represents the output warping path. The aforementioned restrictions are used accordingly to change the output of DTW. For example warping path restriction can be used to force the warping path to progress diagonally by introducing a threshold depth where recursion has to stop, i.e.

$$\gamma(i,j) = \begin{cases} d(T_i, S_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} & i-j| < \delta \\ \infty & \text{otherwise} \end{cases} \qquad 3\text{-}5$$

Where 'δ' represents the threshold value for the maximum depth of the local search function. Figure 3-8 shows the calculated warped path (slope) obtained by the matrix data presented in Figure 3-5. It can be observed that the warping path progresses diagonally from bottom left to top right according to the aforementioned constraints. The DTW has been most commonly used in the literature for the isolated word matching tasks. Prior to introduction of the statistical models (i.e. HMMs) for speech recognition, the DTW has been considered as one of the dominant technique for isolated word recognition (Cassisi 2012). Literature contains a variety of isolated word recognition and ASR approaches that are based on different versions of DTW. The following section presents the latest research work in relation to STD, QbyE, and speech signal matching using a variety of DTW approaches.

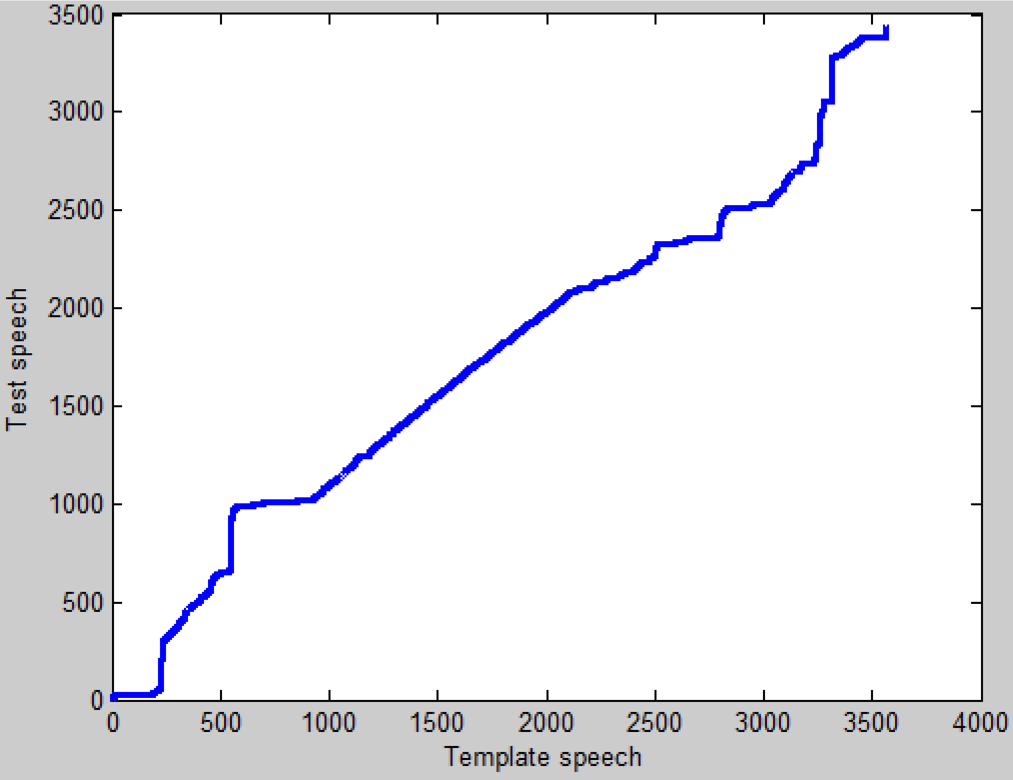


Figure 3-8: Warped Path between Two Speech Utterances

3.5. A Review of Speech Signal Matching

Keyword spotting and speech similarity measurement have been a popular research topic since long period of time. However due to dynamic nature of continuous speech signal in terms of noise interference, time warping, connected speech, and variability in frequency ranges; the tasks of keyword spotting and continuous speech similarity measurement become more challenging. Literature consists of a massive research work related to these tasks that can be categorised into isolated word matching, keyword spotting, and continuous speech matching tasks that are presented in the following sections.

3.5.1. Isolated Word Matching

Isolated word recognition is an easy task as compared to continuous speech matching or keyword spotting due to the discreteness and isolation of the speech signal. Literature contains a high number of research studies where the isolated word matching is addressed using a variety of approaches. An MFCC based isolated word recognition is presented by (Dhingra 2013) where extracted features for test and template word in the form of MFCC are passed to DTW model for the distance calculation. The DTW score for test word is calculated against all template words and the best (minimum) score is picked up as a corresponding best recognized word. Although, this technique works very well (72% accuracy), it is only limited to isolated word/utterance matching. The recognition accuracy will be effected for connected speech.

In the context of speech recognition and signal processing, DTW has been used successfully for time warped speech utterance matching. A signal dependent matching for isolated word recognition is proposed by (Yegnanarayana and Sreekumar 1984) resulting a better performance using FFT for feature extraction and enhanced version of DTW that uses Euclidean distance for signal matching. The algorithm is divided into two stages. The first step includes path construction between template and test (input) word. In the second step, distance between these words is calculated along the path. Matching strategy is based on the Euclidean distance calculated between input and template frames using DTW. The conventional DTW matching algorithms treat same matching for all categories of speech signal. For example, when silence frame is matched with another silence frame, the matching result could be high due to noise in the data. In enhanced version of DTW, weight factor is proposed to differentiate the test signal into silence, voiced, and unvoiced classes which improve the performance of conventional DTW. However, the vocabulary is based on only 6 alphabets that include F, H, L, M, N and X. The enhanced DTW showed improved performance of 78.14% from 56.34% when compared with conventional DTW.

An improved DTW technique is proposed in (Lin and Ji 2010) based on cross correlation for digit recognition. A new approach of slacked start and end point is introduced that improves the performance which is different from traditional DTW algorithms that depends on the performance of end point detection. Cross correlation is used for the similarity measure between reference and template components. The sound was recorded at the sampling rate of 8 KHz, frame size was fixed to 30 milliseconds with 240 sampling points, and overlap was set to 10 milliseconds with 80 sampling points. The system produced improved performance over traditional DTW with limited parameters such as vocabulary size of 10 digits.

In addition to DTW, there are some other methods to deal with the time warped matching of two signals. For example, the landmarks similarity that is based on extracted features such as local maxima, local minima, and inflection points. Similarly, interpolation and extrapolation (down sampling) is another solution to stretch or compress a time warped signal to required length. However; interpolation may lead to change the originality of the function (model fitting) that might be a major cause for mismatching or misidentification in terms of speech signal matching. In proposed research work, a novel technique for dynamic time warping and speech signal tracking is proposed (Chapter 4) that is based on the amalgamation of different speech processing techniques, a dynamic model for object linear motion, and a Kalman filter.

A speaker independent isolated word recognition system is presented by (Peinado 1991) based on HMMs and VQ. The system is trained on a limited vocabulary of 10 Spanish digits spoken by 20 speakers, 3 utterances per speaker for each digit. The training and test data was sampled to 8 KHz and is framed to 256 samples. The system is tested on the same amount of data recorded by same speakers and achieved 90% recognition score. As the system is based on HMM and VQ algorithm, it needs training process on a very specific and limited vocabulary, hence can't be applied for keyword spotting of unlimited vocabulary and language independence.

An isolated word recognizer is introduced by Lipeika (2010) which uses ANN for the training of extracted feature vector of the templates. Power spectral density is calculated by autoregressive method in each frame of speech signal. Autoregressive method is parametric approach for estimation of spectrum of data that can achieve with small data set. After resampling the speech signal to 8 kHz, power spectral density is calculated and forwarded to feature extraction method. Two features; amplitude and location of peaks in each frame were extracted and passed to training process. The system is trained on these two features for available templates of two words '*yes'* and '*no'* using ANN. Only two recordings for each word are used for the training. The system was tested 80 times for both words by the same speaker and 100% accuracy was achieved. Despite of the robust performance of this system, it is limited to recognize only two words. Moreover, it uses a supervised approach (i.e. ANN) for training the network that is not applicable in our proposed system due to unlimited vocabulary.

3.5.2. Keyword Spotting and Continuous Speech Matching

Keyword spotting is based on the partial information extraction (keyword) from a continuous speech signal. Despite of the fact that research has been conducted in the area of keyword spotting since forty years, yet the formulation of the keyword spotting has not been well established (Chen 2014). The related research work can be summarized into three main categories that include Query-by-Example (QbyE) methods, keyword/filler methods, and Large Vocabulary Continuous Speech Recognition (LVCSR) methods. Literature consists a number of keyword spotting approaches in relation to QbyE (Anguera et al. 2014), (Joho and Kishida 2014), (Wang et al. 2011), (Tejedor et al. 2013), (Tejedor et al. 2015), (Abad et al. 2013) and STD (Mandal et al. 2014), (Metze et al. 2012), (Anguera et al. 2013), (Chan and Lee 2013) that use some sort of variations in DTW (Sakoe and Chiba 1978), (Zhang and Glass 2009; 2011), (Chunan and Lin-shan 2010). Over the past decade, most of the related research is focused on novelty of template representation methods (Fousek and Hermansky 2006), (Hazen et al. 2009), (Huijbregts et al. 2011), (Parada et al. 2009), (Wang et al. 2011).

Recent research work introduced in (Abad et al. 2013) addresses the fusion of heterogeneous STD system. In the first step, a number of heuristics are hypothesised for the similarity score estimation and then linear logistic regression method is used for the combination of these scores. The performance is measured using eight different techniques individually as well as by fusing them together using linear regression. It is observed that the STD rate improved with multiple systems fusion as compared to their individual performances. However, this fusion method doesn’t provide the degree of ignorance as introduced in the proposed research work for keyword spotting. Also, the pre-processing in terms of speech enhancement can also improve the keyword detection rate that is addressed in the proposed research. Likewise, (Metze et al. 2012) proposed an effective approach for STD that is based on acoustic segment models. This method amalgamates the self-organising models, query matching, and query modelling processes to construct an efficient STD approach.

An unsupervised spoken term detection using acoustic segment model is presented by (Wang et al. 2011). The aim of the study was to measure the QbyE performance using acoustic segmentation model based posteriorgrams and traditional Gaussian Mixture Model (GMM) posteriorgrams. The acoustic segment models are the unsupervised HMMs of non-transcript speech data. The segmented DTW is applied for the query and test utterance matching and the location of the query utterance are identified. The Fisher and the TIMIT dataset (Garofolo et al. 1993) are used for the experimentation purpose. The system performance is measured using the standard binary classification method. With a tolerance of a single window size in position measurement, a 77.5% for 'recall' is achieved. It is observed that the 'recall' increases to 88.2% with the tolerance of double window size. Despite of the fact that this approach does not uses the speech transcript for the supervised training, it uses the HMMs for posteriorgrams representation that may take huge amount of computational cost.

Spoken term detection in speech for QbyE approach is introduced in (Hazen 2009) for a limited or no in-domain training data. The keyword and template speeches are represented by phonetic posteriorgrams obtained from a phonetic recognition system. The measured posteriorgrams are forwarded to a constrained DTW that measures the warping distance and the position with minimum warping distance is identified as desired keyword. The fisher dataset is used for the performance evaluation while 40 keywords were chosen for the query utterance. It is observed that the average query detection error varies from 10% to 20 % depending upon the DTW constraints. The advantage of this approach is the language independence as there is no training on the phonemes models. However, the accuracy in terms of keyword detection may be a question that is improved in the proposed research work.

Hierarchical posterior based keyword recognition is proposed by (Fousek and Hermansky 2006) where each targeted word is classified by a separate binary classifier against the template utterances. For '*N*' number of keywords, '*N*' parallel binary classifiers are applied that uses the posterior probabilities of the phonemes classes. This technique performs better than other unsupervised keyword spotting techniques in terms of out of vocabulary issue. However, a huge amount of data and time is necessary for this approach that is the major concern addressed in the proposed research study. (Junkawitsch et al. 1996) presented the keyword spotting approach that is based on traditional HMMs and a modified Viterbi algorithm. This approach also suffers from the limited vocabulary issue. Also, it needs a huge training dataset to train the HMMs on the phoneme models.

A keyword spotter is presented in (Bahi and Benati 2009) based on MFCC and energy of the speech signal as feature set. The system is analysed using Hamming window with 36 milliseconds frame length. The VQ algorithm is used for the vector training and codebook generation. Each codebook represents the acoustic features of the signal. The HMMs were used for the probability assignment for observation given a word. If the highest probability is up to a given threshold, the system confirms the detection of the keyword. The system is tested on a limited data of ten speech recordings each having a length of ten seconds. The results show that the system outperforms for this limited dataset with very minor false detection rate. However, this system is based on VQ and HMM that need a high computational time for training learning the models. Moreover, HMM based recognizers depend upon probability assignment that may not be assigned correctly and may cause misrecognition.

One of the most interesting and challenging area of speech pattern matching research is associated with the human-computer interaction based applications. These applications include voice based user interfaces such as voice dialling; call routing, keyword spotting, data entry, and speech to text processing (Hagen et al. 2007). Despite of the significant progress in the performance of ASR since the past decade, there is a limited research focus on the continuous speech tracking and similarity measure. As discussed aforementioned, keyword spotting, spoken term detection, and QbyE research work has been focused over the long period of time. An extension to these approaches may leads to a continuous speech tracking approach.

A speech tracking related approach work is proposed by (Renger et al. 2011) where the children orally read stories to develop/improve the skills of reading and speaking. In such kind of human-computer interactive applications, children are offered the opportunity to orally read the given text and learn about their reading abilities on a real time basis. It can be very helpful in the situation of proof reading for people who want to learn speech contents by heart. Similarly, a very fruitful use of speech tracking is for the identification and localization of word occurrences in a long speech recording. Such kind of application can be used for the security institutions and intelligence agencies.

However; it is not a simple task to develop a robust speech tracking system due to variable length of spoken words, speech continuity and word connectivity, background noise interference, computational complexity, and vocabulary size. Among these issues, the dynamic word length (time warped speech) based on the variable speed of input speech is one of the big challenges to deal with. Different recordings of the same sentence may vary in time durations. In addition, each word within a sentence may also differs in length which may result in the failure of similarity measurement techniques to find the best match in a continuous speech. Consequently, obtaining the accurate information about the test speech frame position with respect to template speech becomes a challenge.

As discussed earlier, most of the related work is based on a variety of DTW approaches which provided the optimal path between time-warped speech signals. However, there are some limitations associated with the DTW approach. For example, the trade-off between computation time and optimal warping path is one of the most common issues associated with DTW (Ratanamahatana and Keogh 2005). This is because in DTW, the computation time is linear to the number of frames (signal length) to be searched through (Cheng-Tao et al. 2014). Extensive efforts were made to enhance the DTW performance in terms of computation time such as segment-based DTW (Chan and Lee 2011), lower-bound estimation for DTW (Zhang and Glass 2011), (Zhang et al. 2012), and a locality sensitive hashing technique for indexing speech frames (Jansen and Durme 2012). Similarly, the conventional DTW approaches use a fixed length frame matching that is problematic in case of speaking rate variations; resulting in poor matching and tracking performance. Also, none of the aforementioned approaches considers the degree of ignorance, model uncertainty, or noise variances that affects the desired performance.

To overcome the dynamics of variable speed of spoken speech, a dynamic state estimator is needed which gives an accurate estimate of the true speech position. Therefore, it is useful to have a robust speech tracking system which is capable of tracking from a speech template and provide an optimal performance in terms of time warped signal matching and localisation from a stream speech. A Kalman filter (KF) is an effective state estimator which can satisfy the criteria of better dealing with variable speed and noise variances. Also, it is an optimal recursive data processing algorithm and has been used for noise cancellation and object tracking for long time (Gelb 1974).

3.6. A Review of Kalman Filter Based Speech Processing

Kalman filter is a recursive solution to the discrete data linear problem that was introduced by R. E. Kalman in 1960. Since that time, research has been conducted on the KF and its applications in different areas. With the advancement in digital computing; the KF has been used in a wide range of applications particularly in the area of target tracking and navigation (Kalman 1960). A KF estimates the state of a process recursively using a set of mathematical equations subject to minimum mean squared error. The most powerful aspect of a KF is the estimations of past, present, and the future states even precise nature of modelled system is unknown as described by *Greg and Gary* (2006). Due to its recursive nature, KF does not require any information on previous states of the model except the last calculated state. It calculates the best optimal state estimation using input measurement and previously calculated state. It has been widely used for dealing with uncertainties by fusing inexact forecasts of a system's state with inexact measurements.

In the past, KF has successfully been applied in sensor fusion, radar tracking, global positioning system, manufacturing, economics, signal processing, and free-way traffic modelling (Grewal and Andrews 2001). A well-known application of state estimate is the tracking of a moving object from radar that has received a great deal of attention in the literature (Daum and Fitzgerald 1983). A discrete time KF is based on mathematical equations in the form of prediction-correction estimator that minimises the error covariance. It addresses the general problem of state estimation of a discrete process that is controlled by linear stochastic difference equation. The implementation details of a KF are presented in (Greg and Gary; Rodman 2006). A practical implementation of KF for a linearly moving object (vehicle) with a uniform velocity and a zero mean random acceleration is presented by (Friedland 1980).

An interesting research work is conducted by (Simon and Jeffery) where KF is utilized for the online filtering application. An extensive use of KF is available for the data fusion in navigation and scene modelling (Joost 2013). Similarly, there is some research work that use KF in the area of image processing and short time non-stationary signal processing. For example, Johnson and Sakaaulis (2003) produced some interesting techniques based on KF for the extraction of components as the trends and seasonal fluctuations in the economic data. Similarly, McGee and Schmidt (1985) used KF as a tool for the aerospace and industry. A detailed study about theoretical and practical aspects of KF is presented by (Gelb 1974). An interesting example of KF application is presented that uses KF for a vehicle navigation to calculate the vehicle position dynamically. Another interesting research work is presented by Ramachandra’s (1988) model that describes the KF based model for a moving vehicle.

In the area of speech signal processing, KF has successfully been used for the speech signal enhancement. More specifically, KF has been used as a noise filter to reduce the noise impact from the speech signal. A KF based speech enhancement algorithm for filtering the speech contaminated by white and coloured noise is presented by (Gibson et al. 1989). By implementing scalar and vector KF, an iterative signal and parameter estimator is presented which can be used for both types of noise. The whole process switches between the corrupted speech measurements given as prior and the estimation of the speech parameters given the enhanced speech waveform. Experiments showed that the algorithm performed better when dealing with coloured noise as compared to white noise. The results were taken on helicopter noise by using state-of-the-art coloured noise assumption in KF which performed very well in terms of signal to noise ratio, sound spectrogram, and output speech quality.

A disadvantage of the aforementioned algorithm is that it does not address the model parameters estimation problem. Also, the performance of this technique is poor against white noise. This problem is solved in research work proposed by (Gannot et al. 1998) where KF is applied for filtering the background noise from actual speech. The expectation-maximisation method is used which estimates the spectral parameters of the speech and noise parameters iteratively. A sparse signal recovery from a series of noisy observations based on KF is presented by (Carmi et al. 2010) in which, a pseudo measurement technique is used to enforce an LP-norm constraint. To retain the linearity of the basic filter, L_1 norm constraint was enforced. The performance results showed that this approach outperformed the Dantzig selector which was considered as an ideal scheme for solving the compressed sensing problem. The distribution of the estimation error over 100 Monte Carlo runs showed that average normalised error for Dantzig selector was 300 as compared to 16 of this approach. Another speech enhancement algorithm based on adaptive KF is presented in (adaptive KF based speech enhancement algorithm) that prevents the explicit estimation of noise and process variances using optimal Kalman gain. The performance of this algorithm is compared with traditional KF based speech enhancement method that uses the voice activity detection for the silence removal. Moreover, this algorithm provides better performance in terms of computation cost because there is no need of filtering step during the optimal Kalman gain estimation.

Despite of the extensive usage of KF in speech related area, there is no clue of its applications towards the continuous speech similarity measurement or speech tracking tasks. In the proposed research work (Khan and Holton 2015), (Khan et al. 2014), for the first time, a KF is used as a feedback system to deal with the time warping effect in the speech signal leading to the introduction of an adaptive frame size.

3.7. Summary

In this chapter, a detailed study on the concepts of distance metrics, DTW, and KF based feedback system is presented. Each method is further explained in depth along with the mathematical formulation and their applications in the literature and in the current research study. The difference between Euclidean distance and DTW is presented. Limitations related to Euclidean distance are discussed and presented in the form of mathematical formulation as well as graphical representation. The constraints associated with DTW approach are addressed. A detailed review of the most recent research work in the area of keyword spotting and continuous speech matching is presented. Similarly, a comprehensive review of the Kalman filter applications in the area of speech enhancement is presented.

4. TIME WARPED CONTINUOUS SPEECH TRACKING AND SIMILARITY MEASUREMENT

4.1. Scope

This chapter consists of four sections, which aim to provide a comprehensive insight into the research contributions in terms of time warped continuous speech tracking (TWCST) and similarity measurement approach. Section 4.2 provides an introduction to the problem followed by a brief discussion on Kalman filter (KF) based object localisation. Section 3 addresses a detailed discussion on the formulation of a dynamic state model for TWCST approach. The concept of adaptive framing, search region, pitch tracking based silence removal, and tuning of the KF are explained. Sequential components of TWCST that include speech enhancement, spectral analysis, feature extraction, and similarity measure are presented. In Section 4.4, detailed discussion on evaluation methodology, simulation settings, simulation tools, and speech corpuses is presented. In Section 4.5, the experimental results and performance evaluation is presented using statistical metrics and validation methods for binary classification and speech tracking. Results for the proposed TWCST approach are compared with the existing DTW based methods and presented in the form of statistics and graphical representation. Finally, a summary of the chapter is presented.

4.2. Introduction

Speech signal based pattern matching and similarity measure have been a challenging research area since 1950's. For past three decades, steady improvements have been shown in speech signal based pattern matching. More difficult tasks have been considered with the passage of time and different techniques have been applied to achieve good performance (Arora and Singh 2012, Hagen et al. 2007). Time warped speech tracking and signal matching is an interesting and challenging research topic that is introduced in the proposed research work (Khan and Holton 2015). The practical implementation of speech tracking may face the speech dynamics issues. In this sense, time warping is one of the challenging issues to be focused that are resolved by innovative approach of adaptive framing (Khan and Holton 2015), (Khan et al. 2014) in the research work presented in this chapter.

To overcome the dynamics of background noise and variable speed during speech tracking, a dynamic state estimator is needed which gives an accurate estimate of the true speech position. A KF is one of the effective state estimators that can satisfy the criteria of better dealing with variable speed and noise. Also, it is an optimal recursive data processing algorithm and has been used for noise cancellation and object tracking for long time (Gelb 1974). The deployment of KF for speech tracking is a novel approach that is introduced very first time in the proposed research (Khan et al. 2014). A Dynamic State Model (DSM) is presented based on equations of linear motion. In DSM, initially, a fixed length frame of test speech is considered as a unidirectional moving object by proceeding it frame by frame along the template signal. The position estimate in template speech for corresponding test frame at current time is calculated using equations of motion. Real time applications of speech tracking may be proof reading, learning speech contents by heart, and content matching for copyright checks. Similarly a productive use of speech tracking is the identification and localization of word occurrences in a long speech recording also called keyword spotting. Such kind of application may be useful for the intelligence agencies.

4.3. Proposed Method for Time Warped Continuous Speech Tracking Using Kalman Filter and Mel Frequency Cepstral Coefficients

The research contribution presented in this scenario introduces a time warped speech tracking approach based on the amalgamation of sequential and parallel processes to deal with variable speed of speaker's speech, silence removal and background noise reduction (Khan and Holton 2015). First, speech signals are passed to pre-treatment process for speech quality enhancement in term of silence removal and noise reduction. Second, the framing process is applied to the enhanced signal that divides the entire test signal into fixed length frames. All frames are then passed sequentially for spectral analysis which uses STFT for the time to frequency domain conversion. The STFT based spectrogram is useful for extraction of features that don't exist in time domain.

The most dominant features are extracted from frequency domain frames which are known as MFFC and passed to a distance metric. Euclidean distance is then used for the calculation of similarity scores between features of test frame and all overlapped template frames in a specified search space. Position of the best matched template frame is selected as the current location for corresponding test frame. Simultaneously, a DSM provides the predicted position of the same test and template frames using equation of linear motion. Both; position estimate by DSM and distance metric are forwarded to KF along with the noise variances and the best estimated frame position in the template speech for the current state at time '*t'* is calculated. Finally, forecasting of the noise variances, template frame size, and search region limits are updated according to the KF output.

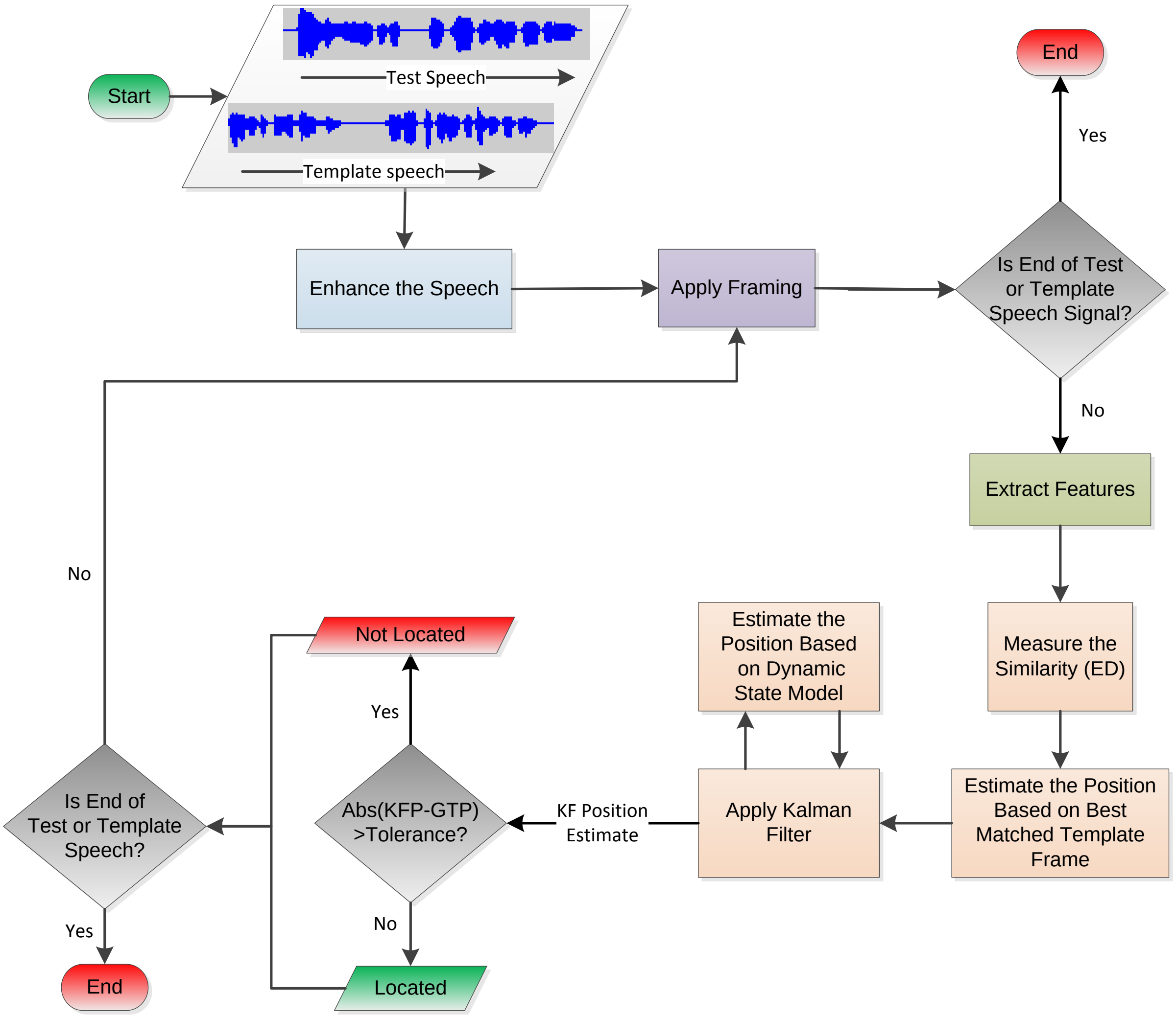

Start
Test Speech
Template speech
Enhance the Speech
Apply Framing
Is End of Test or Template Speech Signal?
Yes
End
No
Extract Features
Measure the Similarity (ED)
Estimate the Position Based on Best Matched Template Frame
Apply Kalman Filter
Estimate the Position Based on Dynamic State Model
KF Position Estimate
Abs(KFP-GTP) >Tolerance?
Yes
Not Located
No
Located
Is End of Test or Template Speech?
No
Yes
End

Figure 4-1: Flowchart for Time Warped Continuous Speech Tracking Approach

The entire process runs recursively to provide the best matching position of the test speech frame in the template at each time step. A sequential flowchart for TWCST approach is shown in Figure 4-1 followed by the technical details for each component.

4.3.1. Pre-Processing

Most of the time, speech signal consists of silence parts that may be a major cause of mismatching and misidentification. In the proposed approach, the pre-processing is used for the enhancement of input speech signal quality in terms of sampling and silence removal. The human vocal sound frequency range varies significantly from one person to another. However, the frequency range of 300Hz to 3.4 kHz is found the best frequency range for the speech intelligibility and speaker recognition (Nortel 2002). According to the Nyquist sampling theorem, phenomenon of aliasing can be prevented if bandwidth of a sampled signal is equal to half of the sampling frequency of that signal. Hence, following these facts and the conducted research experiments, the sampling frequency is set to 8 kHz in the current research.

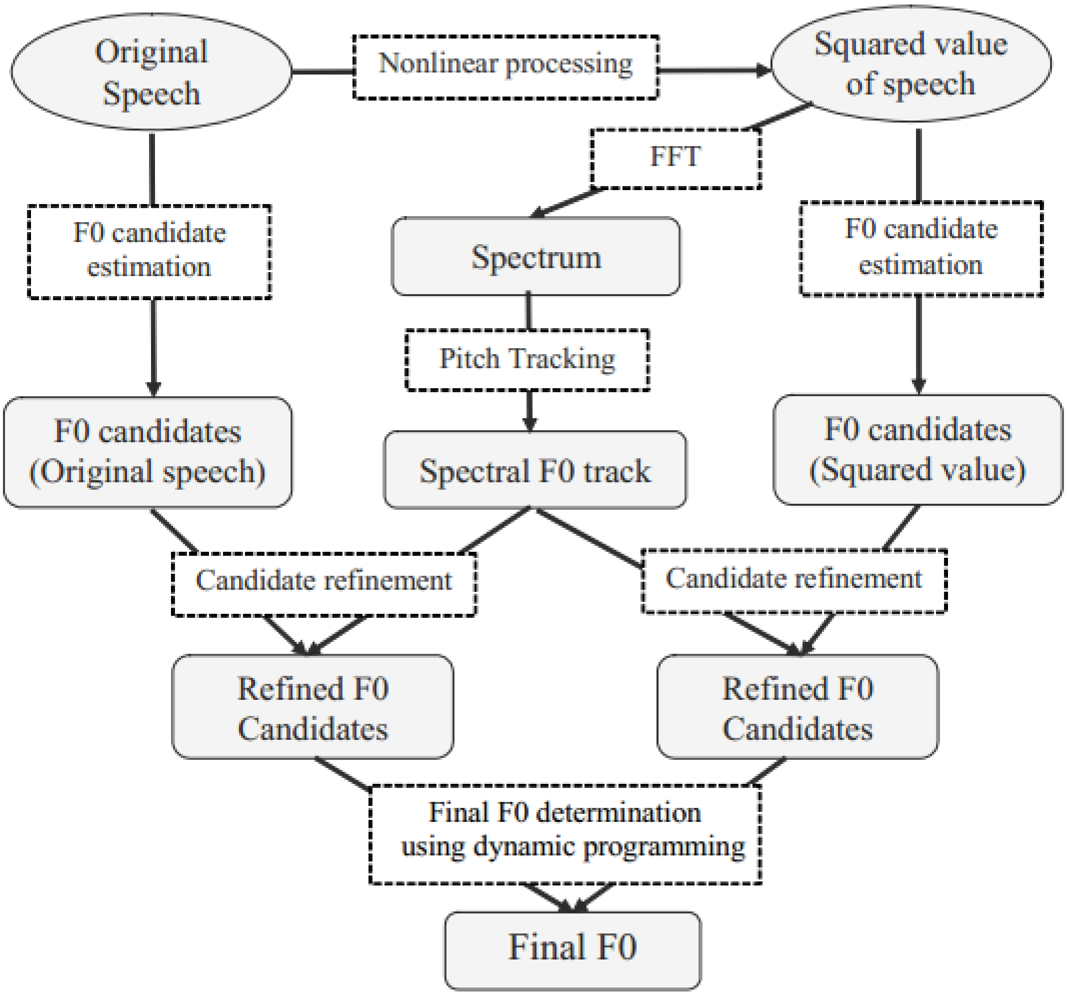


Figure 4-2: Implementation Design of YAAPT (Zahorian and Hu 2008).

There are a number of techniques in the literature that uses time and frequency domain features (e.g. energy, zero cross rate, spectral centroid) to remove the silence part of speech signal (Section 2.2.1.4. for detailed review). For example (Sharma and Rajpoot 2013) proposed a silence removal technique based on energy and zero cross rate (ZCR). Similarly, a voice activity detection technique is proposed in (Giannakopoulos 2014) based on zero cross rate and spectral centroid. In the proposed research, a robust pitch tracking approach (YAAPT) is deployed that was proposed by Zahorian and Hu (2008). As compared to conventional pitch estimation approaches, it estimates the F_0 using multiple information sources that are based on time and frequency domain features. A block diagram of YAAPT is presented in Figure 4-2.

It can be observed that fundamental frequency is estimated using multiple information sources. Firstly, the original speech and squared speech signals are cross-correlated to extract the peaks. Meanwhile, a smoothed pitch track is obtained from the spectral information using FFT. Information by the aforementioned sources is combined using dynamic programming to extract the fundamental frequency estimate. As the F_0 components do not exist in the silence part of speech (Sharma and Rajpoot 2013), we can easily eliminate the silence frames after the F_0 estimation. The Original speech signal is forwarded to the pitch tracking algorithm which searches for the existence of fundamental frequency components in each frame of input speech signal. All frames having the fundamental frequency components are produced as an output. Finally, these frames are recombined and the silence free speech signal is reconstructed which is used for further processing. Figure 4-3 demonstrates the steps for the utilization of pitch tracking and reconstruction of silence free speech signal.

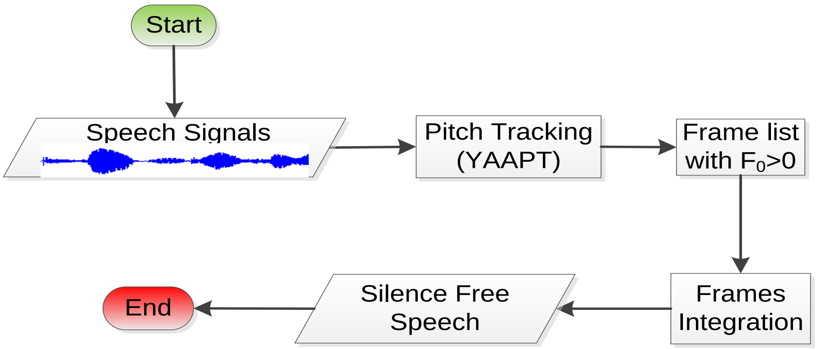


Figure 4-3: Silence Removal from Speech Signal

In the current research study, the aforementioned silence removal techniques are compared based on their performances for the TWCST (Appendix B, Figure 4-16). Figure 4-4(b) demonstrates a test case for silence removal performance using energy, ZCR, and F_0. It can be analyse that even most of the silence segments are identified yet, the pitch tracking based silence removal produces more efficient results in terms of discreteness while applied on the same speech signal as shown in Figure 4-4(c). More information about the silence removal approaches can be found section 2.2.1.4.

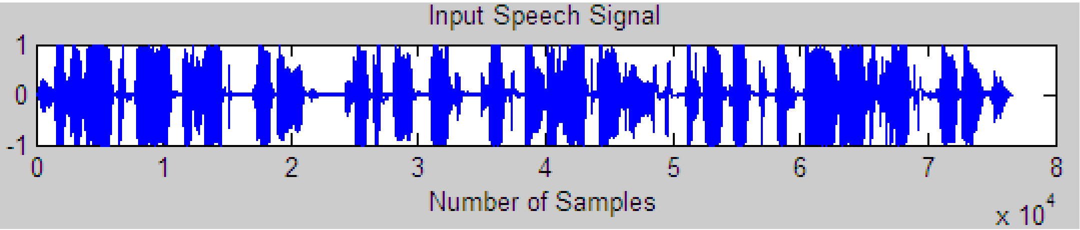


Figure 4-4 (a): Input speech signal with silence segments

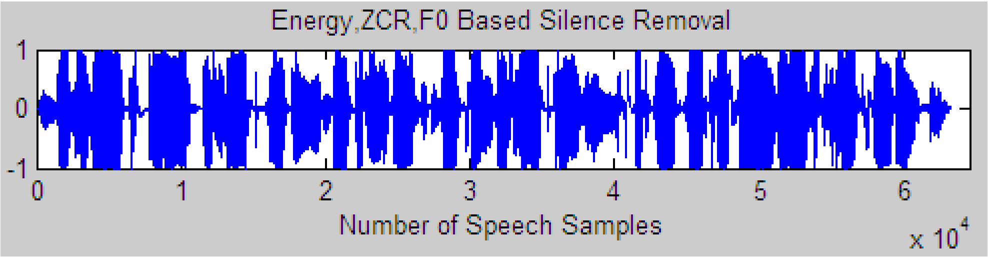


Figure 4-4 (b): Silence Removal Based On Energy, ZCR and F0

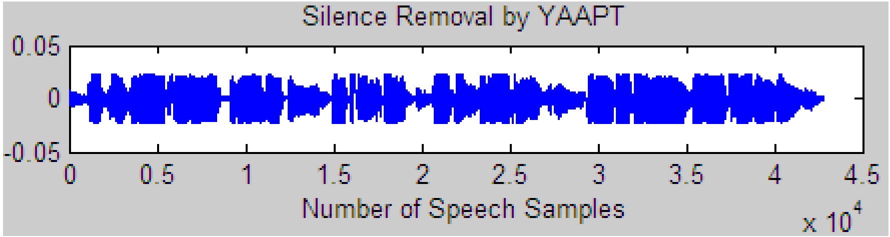


Figure 4-4 (c): Silence Removal Performance by YAAPT

To summarise, the silence free speech is forwarded for the segmentation process. The test speech signal is segmented by a fixed frame size while the template frames size changes dynamically depending upon the varying speaking speed. Speech framing is a process of decomposing the speech signal into smaller units. Because of the slowly varying nature of the speech signal, it is common to process speech in blocks (also called “frames”) of 10 to 50 milliseconds, over which the speech waveform can be assumed as a stationary signal (Ravindran et al. 2010). In the proposed research, the speech signal is decomposed into frames of 40 milliseconds duration and forwarded for further processing. A detailed discussion on framing process in presented in Section 2.2.1.1.

4.3.2. Spectral Analysis and Feature Extraction

Generally, human voice contains important information such as gender, emotion and identity of speaker that can be categorized in different classes. To extract this information, speech signal is needed to be analysed in frequency domain. Spectral analysis is the component of speech tracking system that converts the time varying speech signal into frequency domain. One of the most common frequency domain transformations is known as FFT or the more efficient STFT. The STFT is used most of the time when a signal is time varying or when it is desirable to have better time localization as in our case of speech tracking. A mathematical implementation of STFT is presented in section 2.2.2. Feature extraction component calculates the dominant features out of time and frequency domain signals that may provide sufficient information to represent the speech signal. A speech signal harbours very rich information which can easily classify the words, speaker, gender, and emotion etc., if no useful information is lost during the feature extraction. The MFCCs features have been successfully used in ASR areas and considered as the most dominant and distinguishing features of human speech (Dhingra et al. 2013), (Dave 2013). A detailed discussion on MFCC features, their mathematical formulation and applications in the related areas is presented in Chapter 2, Section 2.2.3. Figure 4-5 presents a block diagram for extracting the MFCC features from a speech signal.

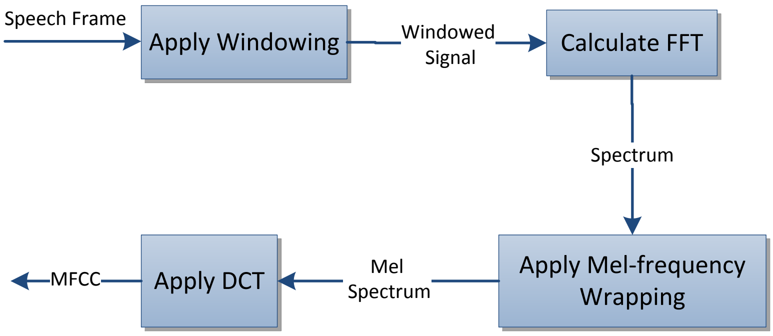


Figure 4-5: Block Diagram for MFCC Features Extraction

4.3.3. Similarity Measurement

In the next step, the normalised extracted MFFC features for both; test and template frame are forwarded to measure the similarity score. Euclidean distance is used as a similarity metric. As Euclidean distance provides dissimilarity score, fewer score means more similar. A detailed formulation and related work for Euclidean distance is presented in Chapter 3, Section 3.3. Suppose, the mean MFCC feature vectors for test and template speech frame with are represented as $km = \{km_{1,} km_{2,} ....km_{n}\}$ and $tm = \{tm_{1,} tm_{2,} ....tm_{n}\}$ respectively. The normalised Euclidean based similarity measure for aforementioned features is then measured by:

$$S_m = 1 - \sum_{i=1}^{n} \sqrt{\left(\frac{km_i}{|km|} - \frac{tm_i}{|tm|} \right)^2} \qquad 4\text{-}1$$

Output measurement S_m gives the similarity measurement between test and template speech features which is normalised to get the similarity measurement values within the ranges of 0 and 1 for each pair of test and template speech frame. These measurements are used as the match/mismatch beliefs that are forwarded to further processing.

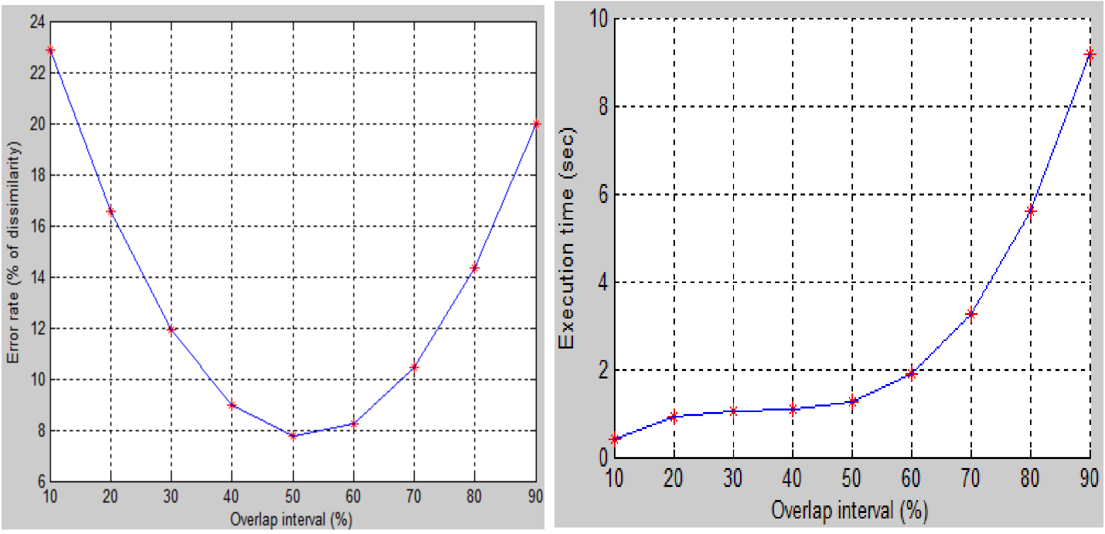
The process of feature extraction and similarity score calculation for the template frames iteratively runs until the end of a *search region*. A search region is a segment of the template speech signal that starts at synchronised position with test frame and ends a frame length after the test frame position as shown in Figure 4-9. The test frame proceeds along the search region and the likelihood for the each overlapped frame is calculated in terms of the probability distribution over the specified space $'2f'$ of search region given as prior information, where $'f'$ is the frame size in template speech. Thus the aim is to calculate the maximum likelihood estimate (MLE) of the position from the calculated similarity distribution. For each frame shift f_i (i = 1, 2 ...n and n = number of shifts), the likelihood positions $f(l_i)$ are calculated using:

$$f(l_i) = l_{cur} + i(f - \Omega) \qquad 4\text{-}2$$

where;

$$l \in \left\{ \begin{array}{ll} [l_{cur} - f/2, l_{cur} + f/2] & \text{if} \;\; l_{cur} \geq 1, \\ & \;\;\;\;\; l_{cur} + f \leq L \\ & \\ [l_{cur} - f/2, l_{cur}] & \;\;\; \text{if} \;\; l_{cur} + f > L \end{array} \right\}$$

In equation 4-2, f_i is the i^{th} overlap position of template frame corresponding to the current position of the test frame and l_{cur} is the current position (i.e. position before the MLE calculation) of test frame in the search region. The 'Ω' is the overlap interval and varied to check the performance and '*L*' is the total length of the search region. It is experimented that the best similarity performance is achieved with the Ω=*f*/2 which also supports the experiments conducted by Bob (2009). Figure 4-6(b) indicates that the execution time is directly proportional to Ω. It can be easily observed in Figure 4-6 that the execution time is almost stable (slightly increasing) when Ω is within the range of 1% to 50%.



(a) Error Rate (dissimilarity) **(b)** Execution Time

Figure 4-6: Template Frame Overlap Effects On Error Rate and Execution Time Performance

In order to obtain the normalised distribution over the specified range, the probability distribution is normalised using:

$$\Phi_l = \frac{\Phi_{l_i}}{\sum_{i=1}^{n} \Phi_{l_i}} \qquad 4\text{-}3$$

where;

Φl is a vector of the similarities calculated by a feature based distance metric when frame overlap is also taken into account*; i = 1…n* represents the number of elements in Φ_l.

$$l_{est} = MLE = \arg\max_{i} \Phi_{l_i} \quad \text{4-4}$$

The best estimate of the position with regard to Φ_l is calculated by the above equation which provides the maximum likelihood estimate (MLE) of the best matched position of the template frame in the search region for corresponding test frame. This MLE is forwarded to KF model as the observation at current time.

4.3.4. Mathematical Formulation of the Kalman Filter and Dynamic State Model

Generally, KF is a recursive process which provides an optimal estimate by taking dynamic model and all possible observed values into account as inputs. The prediction step in KF calculates the position estimate of the template frame corresponding to test frame on the basis of a DSM. In DSM, the test speech signal is segmented into fixed size frame where each frame is considered as a unidirectional moving object by preceding it within the search region of template signal. The DSM estimated position and position measured by the feature based Euclidean distance are passed to the update step. The update step takes both position estimates as control inputs along with process and measurement noise variances to provide a best position estimate. Mathematical formulation of KF (Figure 4-7) for the proposed speech tracking approach is explained below.

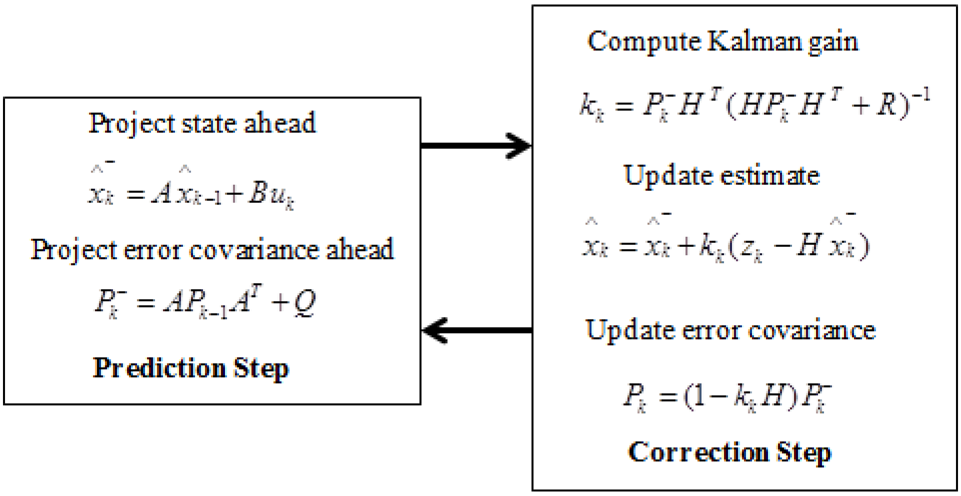

Project state ahead
$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k$
Project error covariance ahead
$P_k^- = AP_{k-1}A^T + Q$
Prediction Step
Compute Kalman gain
$k_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$
Update estimate
$\hat{x}_k = \hat{x}_k^- + k_k(z_k - H\hat{x}_k^-)$
Update error covariance
$P_k = (1 - k_k H)P_k^-$
Correction Step

Figure 4-7: Kalman Filter Prediction-Correction Steps (Rodman 2006)

a) Prediction Step and Dynamic State Model

In order to overcome the position noise using KF, the process that is being measured must be able to be described by a linear system. The equations for speech frame position prediction and correction are modelled by a linear system as follows:

$$x_{t+1} = Ax_t + Bu_t + f_t \qquad 4\text{-}5$$

$$z_t = Hx_t + g_t \qquad 4\text{-}6$$

where, A, B and H are state, input, and output matrices respectively, '*t*' is the time index, 'x' is the state of the system at time '*t*', 'u' is known input to the system which is $\Delta\dot{f}$ in our case, z_t is the measured output, f_t and g_t are process and measurement noise respectively. In the context of speech tracking, the noise is taken into account in terms of variable length of same spoken words and hence short phrases, paragraphs, and long speech recordings. In speech tracking system, a DSM is implemented by considering the test speech signal as a moving object along the template speech frame by frame with the time progression. For test frame, suppose

p_t is position, v_t is initial velocity, v_{t+1} is final velocity, and $\Delta\dot{f}$ as change in the template frame size at time '*t*'. The equation for current velocity will be:

$$v_{t+1} = v_t + \Delta\, \dot{f}\, T \qquad \text{4-7}$$

At very first step, velocity is set to sample rate (8000 samples/sec) and recursively updated with respect to $\Delta\dot{f}$. There is no change in the frame size of template speech at initial step, i.e. $\Delta\dot{f} = 0$ which is updated after each iteration using Equation 4-14 and passed to the system to update the search region and adapt the template frame size accordingly. The equation for position '*p*', at time '*t*' can be represented as:

$$p_{t+1} = p_t + Tv_t + \frac{1}{2}\Delta\, \dot{f}\, T^2 \qquad \text{4-8}$$

Where; p_{t+1} is the current test frame position at time $t+1$ and p_t is the previous test frame position at time '*t*'. '*T*' is the fixed time interval (40 milliseconds) which represents the fixed length of test speech frame. Initially, starting position p_t is zero and end position p_{t+1} = 320 samples (test frame size). This implies that the change in template frame size $\Delta\dot{f}$ shows the relative change in speed of test speech. A state vector '$\varkappa$' can be defined consisting of position and velocity as components.

$$x_t = \begin{bmatrix} p_t \\ v_t \end{bmatrix}$$

By comparing Equation 4-5, 4-6, 4-7, and 4-8, the linear system can be written as:

$$x_{t+1} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} x_t + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \Delta\, \dot{f} + f_t \qquad \text{4-9}$$

$$z_t = \begin{bmatrix} 1 & 0 \end{bmatrix} x_t + g_t \qquad \text{4-10}$$

To control the test speech signal with some sort of feedback system, an accurate estimate of test speech frame position and speed would be needed. In other words, an optimal estimate of the state is needed by considering various noise sources of a speech. This is where KF comes in. In the above described system, f_t is the process noise and g_t is the measurement noise. KF solution does not apply until certain assumptions about f_t and g_t are made. In the proposed system for speech tracking, it is assumed that there is no correlation between f_t and g_t that means both are independent random variables. Based on experimental results for KF tuning addressed in next section, the initial values for f_t and g_t are set to 0.72 and 0.28 respectively. The covariance matrices for process and measurement noise can be defined as:

$$Q = E(f_t f_t^T)$$

$$= E\left(\begin{bmatrix} \mathrm{p} & \mathrm{v} \end{bmatrix}\begin{bmatrix} p \\ v \end{bmatrix}\right) = E\left(\begin{bmatrix} p^2 & pv \\ vp & v^2 \end{bmatrix}\right) = f_t \times \begin{bmatrix} \frac{T^4}{4} & \frac{T^3}{2} \\ \frac{T^3}{2} & \mathrm{T} \end{bmatrix} \qquad \text{4-11}$$

The above matrix represents the standard deviation in position and velocity.

$$R = E(g_t g_t^T) = E(g_t^2) \qquad \text{4-12}$$

This represents the standard deviation of measurement noise covariance.

b) Update (Correction) Step

The predicted test frame position by DSM along with the estimated position by distance metric at time '*t*'; are passed to update step of KF. This step is also called correction step and it updates the estimate with the updated Kalman gain 'k', process noise variance, measurement noise variance, and state estimate 'x '. Equation for 'k' is represented as:

$$k_t = P_t^- H^T (H P_t^- H^T + R)^{-1} \qquad \text{4-13}$$

Where

$P_t^- = AP_{t-1}A^T + Q$; $P_t = (I - k_t H)P_t^-$,

$A = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$, $H = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $Q = E(f_t * f_t^T)$, $R = E(g_t^{\ 2})$

Equations for state estimate and error covariance update can be represented as:

$$\hat{x}_t = \hat{x}_t^- + k_t(z_t - H\hat{x}_t^-) \qquad 4\text{-}14$$

$$P_t = (I - k_t H)P_t^- \qquad 4\text{-}15$$

The update step consists of the above three Equations (4-13, 4-14, and 4-15) which involve matrix manipulations. The '-1' superscription in Equation 4-13 indicates the matrix inverse and 'T' superscription represents the transpose of the matrix, 'z_t' in Equation 4-14 represents the measurement provided by feature based Euclidean distance at current time 't' and $\hat{x}_t$, $\hat{x}_t^-$ represents the state estimate provided by DSM at time $t, t-1$ respectively. The estimated error covariance is represented by 'P' which is the estimate of initial test frame position variance and is equal to process noise covariance matrix 'Q' at the first step (i.e. P=Q). The equation for 'k' shows the state estimation dependency upon process and measurement noise covariance. The above set-up represents the cycle of KF for speech tracking using a feature based similarity measurement and DSM position estimates as input parameters.

4.3.5. Tuning the Kalman Filter

Practically, the measurement noise covariance 'R' can be measured prior to filter operation because during the filter operation, as the process is needed to be measured, the process noise covariance 'Q' measurement can be a challenge. Generally speaking, a robust filter performance may be achieved by tuning the filter parameters 'Q' and 'R'. Most of the time; the tuning is performed offline, frequently with the help of another (distinct) KF in a process generally referred to a system identification (Greg and Gary 2006). It can be analysed from measurement update that in case of constant values of '*Q*' and '*R*', estimation error covariance P_k and Kalman gain k_k will stabilize quickly and then remain constant. This means that these parameters can be computed prior to filter operation by offline tuning or by determining the steady state values (Grewal and Andrews 1993). Variations in the values of '*Q*' and '*R*' indicate the dependency (level of trust) of the system. Greater value for a variance means less dependency on the corresponding measure and vice versa.

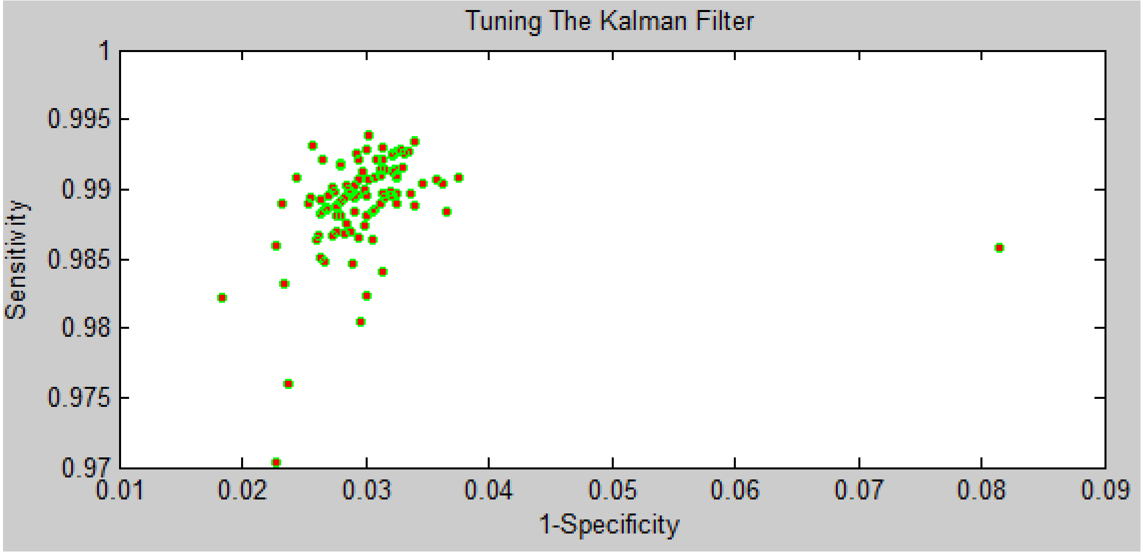
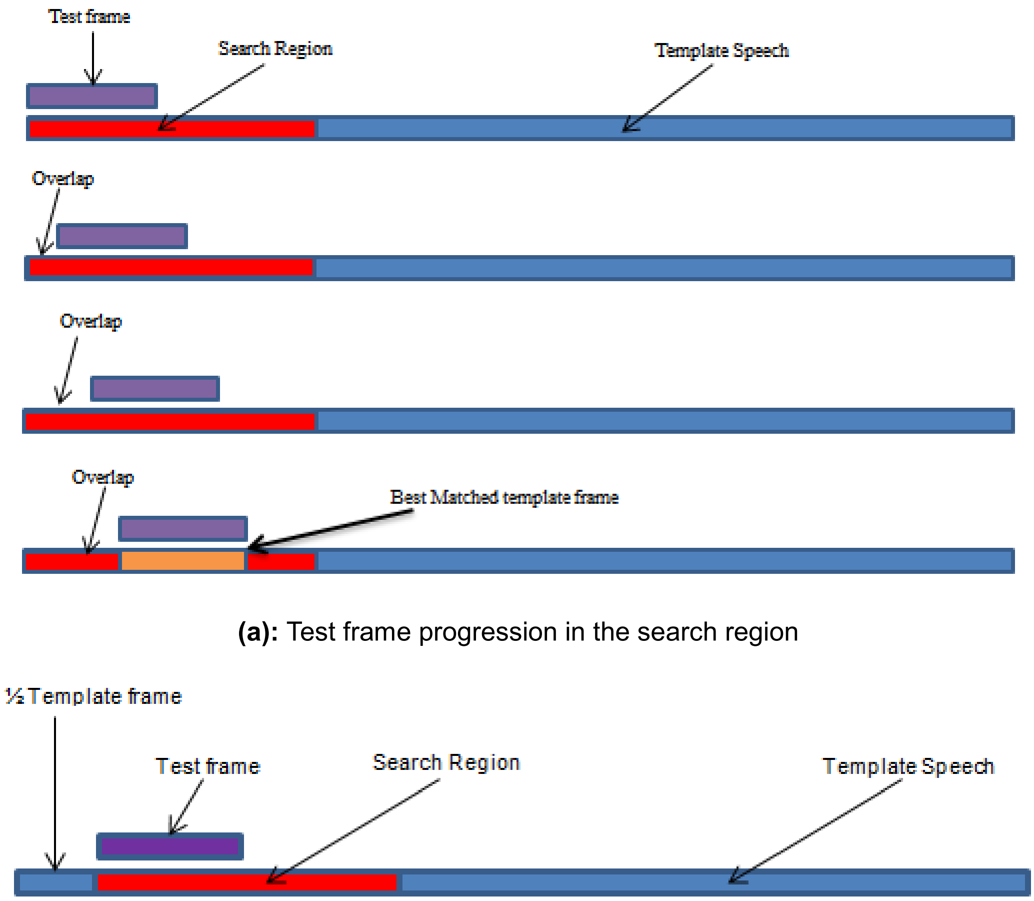


Figure 4-8: KF Noise Variance Selection

In TWCST approach, values for measurement and process variances are validated using the ROC curve points that are retrieved by varying them from 0 to 1 with a lag of 0.01 as shown in Figure 4-8. The entire setup is tested on a speech dataset acquired to conduct the case study (Table 4-3) and the best values for process and measurement noise variances are selected based on sensitivity and specificity.

4.3.6. Frame Size and Search Region Adaptation

A major issue associated with existing time warped speech signal matching approaches is the dynamic length of speech. Same word recorded at different time might have variable length due to the dynamic spoken speed of speaker. To overcome the time warping issue, an efficient approach is introduced based on dynamic adaptation of frame size and search region with the progression of time. Initially, both test and template signals are divided into segments or frames having the same length. The figure below shows the template frames and search region updating process along the time progression.



(a): Test frame progression in the search region

(b): Updated Positions of Search Region, Test and Template Frames

Figure 4-9: Adaptive framing and progression of test frame along the template speech

Suppose, $\dot{f}_1$ is the fixed frame size for the test frame and $\dot{f}_2$ is the initial frame size for the template frame. At very first step, $\dot{f}_2$ is set equal to $\dot{f}_1$ and is recursively updated according to KF estimates thereafter as shown in Figure 4-9(b). Suppose $\Delta \dot{f}$ is the change in template frame size based on the current velocity, the template frame size $\dot{f}_{2(t+1)}$for the next iteration can be measured by adding $\Delta \dot{f}$ to $\dot{f}_{2(t)}$. This implies that the adaptive frame size for template signal depends upon the speed of test speech and changes with respect to the KF position estimation. The relationship of KF position estimate and adaptive frame size can be represented as:

$$\Delta \dot{f} = \frac{v_{t+1} - v_t}{T} \qquad \text{4-16}$$

$$\dot{f}_{2(t+1)} = \dot{f}_{2(t)} + \Delta \dot{f} \qquad \text{4-17}$$

The above equations calculates the template frame size for the next iteration (i.e. at *t*+1) which will be either decreased or increased by the amount '$\Delta \dot{f}$'. In other words, '$\Delta \dot{f}$' represent the acceleration/deceleration of the template speech frame due to the phenomenon of time warping. The whole process runs recursively and the best estimated frame position by a KF and the calculated $\dot{f}_{2(t+1)}$ are forwarded to framing process. Figure 4-10 represents the output for frame size adaptation for a test case. It can be analysed that the template frame size changes instantly using Equation 4-17 and reflects the natural time warping phenomenon in a speech signal.

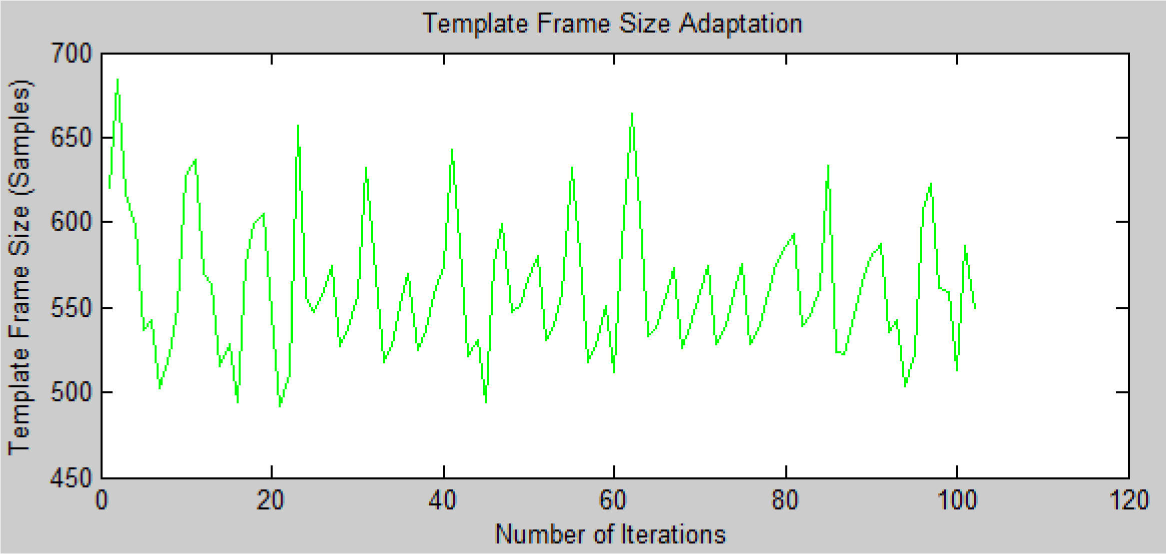


Figure 4-10: Adaptive Frame Size in Template Speech Signal

4.4. Experimental Setup

To achieve the satisfactory performance, a number of factors are set by iteratively analysing the experimental results and updating the setup values. These factors consist of KF tuning, recording devices, processing tools, processing devices, and recording environments etc. For recording purpose, the SENNHEISER e935 is used which is a vocal dynamic microphone that consists a built in noise filter. Also, the dataset is recorded in a noise free research lab environment. Table 4-1 shows the simulation settings for the experimentation of speech tracking performance.

Table 4-1: Simulation Settings

<table>
<tr><th colspan="2">Hardware specifications</th></tr>
<tr><td colspan="2"><u>Processor:</u> Intel® Core™ i5 CPU
<u>Installed memory:</u> 4 GB
<u>System type:</u> 32 bit Operating system
<u>OS:</u> Window 7 Home Premium
<u>H-Disk:</u> 500 GB
<u>Microphone:</u> SENNHEISER e935</td></tr>
<tr><td><u>Simulation tools</u></td><td>Matlab R2009a, PRAAT, SFS, Audacity</td></tr>
<tr><td><u>Sampling frequency</u></td><td>8000 Hz</td></tr>
</table>

Initial frame size	320 samples
Overlap amount	50%
Search Region	2 * frame size = 640 samples
Tolerance	½ template frame
Kalman filter variables	
T	0.04 seconds
p_t	0
p_{t+1}	320 (40 milliseconds)
Δf	0 (changes dynamically)
f_t	0.72
g_t	0.28

4.4.1. Evaluation Methodology

The evaluation methodology entails experiments for multiple settings that involve KF variables setting, silence removal approaches, similarity metrics, search space constraints, and feature selection. A number of metrics have been used in the literature for the validation of keyword spotting approaches. However, the most relevant are the gold standards used for the performance evaluation of a binary classifier (Soluade 2010). This is because the output of test and template speech frames is in the binary form (i.e. match or mismatch). Table 4-2 presents the detailed metrics that are used for the validation of the proposed TWCST approach.

True Positives (TP): All instances where the recognizer correctly measures the template frames locations for corresponding test speech frames while considering the tolerance level with respect to start and the end of a frame position.

True Negatives (TN): The recognizer correctly rejects the out-of-grammar utterances.

Table 4-2: Results Validation Metrics

	Condition as determined by Gold Standard	
Total Population	Condition Positive	Condition negative
Positive Match	**True Positive**	**False Positive**
Negative Match	**False negative**	**True Negative**
Accuracy (ACC) = (Σ True positive + Σ True negative) / Σ Total Population	True positive rate (TPR), Sensitivity, Recall = Σ True positive/ Σ Condition Positive	(Type I Error) False positive rate (FPR), Fall-out = Σ False positive/ Σ Condition Negative
	(Type II Error) False negative rate (FNR) = Σ False negative/ Σ Condition Positive	True negative rate (TNR), Specificity (SPC) = Σ True negative/ Σ Condition Negative
Precision = Σ True positive/ Σ Positive Match F1 Score = 2x Precision x Recall/ (Precision + Recall)	Positive likelihood ratio (LR+) = TPR/FPR	Negative likelihood ratio (LR−) = FNR/TNR

False Positives (FP): Represents all instances that are recognised as best matched template frames locations when there is no similarity exists between test and template frames.

False Negatives (FN): The recognizer incorrectly rejects the in-grammar utterances.

Sensitivity (Recall or True Positive Rate): Probability that a test frame is positively matched with template frame with high confidence when the test frame is actually in-grammar (template frame). This is expressed as a percentage of all the in-grammar matches.

Specificity (True Negative Rate): Probability that a test speech frame is matched as out-of-grammar when it is indeed out-of-grammar and is therefore not accepted by the recognizer. This is expressed as a percentage of the all out-of-grammar matches.

Accuracy: This is a percentage of all the matches that were correctly classified.

Positive Likelihood Ratio: The ratio of the probability of positively recognizing a test frame with high confidence when an in-grammar test frame is spoken and the probability of positively recognizing a test frame with high confidence when an out-of-grammar test frame is spoken. This is basically the True Positive Rate/False Positive Rate.

Negative Likelihood Ratio: The ratio of the probability of rejecting an in-grammar test frame and the probability of rejecting an out-of-grammar test frame.

Precision (Positive Predictive Value): Fraction of retrieved instances that are relevant.

F-Score: Also known as F1-Score that represents the weighted average of precision and recall with best output value at 1 and worst at 0. It provides a good indication of system accuracy while considering simultaneously recall and precision with varying weights.

Receiver Operating Characteristics (ROC): In binary classification, the ROC curve is a standard indication for the trade-off between sensitivity and specificity as its discrimination threshold varies (Soluade 2010). This curve is achieved by plotting the true positive rate against the false positive rates for a varying threshold. For the proposed TWCST approach, this threshold defines the decision boundary for the similarity metric to decide whether the test and template frame are matched or mismatched. Thus the ROC curve provides a tool to select the optimal approach with the best threshold value and discard suboptimal thresholds. Closer the ROC curve to the upper-left corner indicates the accuracy rate of the test. Similarly, area under the curve (AUC) is also considered as a standard measure for the ROC curve analysis. The AUC with a value of 1 represents a perfect test; 0.5 represents a worthless (random) test and 0 represents a failed test.

Tracking Accuracy: Likewise the matching accuracy; the tracking accuracy indicates the perfection level for the test frames localisation in the template speech. However, tracking accuracy can’t be measured using the sensitivity and specificity. In TWCST, the tracking accuracy indicates the percentage of correct position estimation of test speech frames in the template speech and is calculated by:

$$trackingAccuracy = \frac{\sum_{i=1}^{N} (KF\_Estimate)_i}{N} \quad \forall \quad \left|KF\_Estimate - groundTruth\right| < tolerance \qquad 4\text{-}18$$

Where, ‘N’ represents the total number of iterations and ‘$groundTruth$’ are the actual positions in the template speech for corresponding test frames.

4.4.2. Simulation Tools

A number of simulation tools have been deployed based on required functionalities. Matlab R2009a is used as a main tool for the experimentation, speech analysis and graphical representation. Matlab is considered as a special simulation tool for the signal processing. One of the main advantages of using Matlab is the availability of huge built-in libraries. Different toolboxes (e.g. Voice-box, Phase-Vocoder) are available that can easily be utilized for speech manipulation, feature extraction, distance metrics, KF, and graphical interfaces.

Audacity 2.0.5 is used for the speech recordings and manipulation. Audacity is an open source cross-platform multi-track audio editor that is normally used for recording and editing sounds. It may be utilized for speech editing in terms of sample rate conversion, noise reduction, silence removal, speech effects, tempo editing, and channel manipulation. Also, it provides the facility of simultaneous manipulation of multiple speech files and broadcasts to output file names in a sequence. Another well-known open source platform (PRAAT) is used for speech annotations. The PRAAT also provides the facility of analysing different speech features that include; formant frequencies, pitch tracking, speech intensity, and spectrogram. Moreover, it provides a user friendly interface for graphical representation of these features. Similarly, Speech Filing System (SFS) has been used for speech processing and feature analysis. The SFS provides a huge functionality set in form of executable that can be export easily for the reusability purpose. The snapshots of aforementioned functionalities by these simulation tools are presented in Appendix C.

4.4.3. Speech Corpus and Dataset

The research experiments are conducted on various open source and proprietary speech dataset consisting speech recorded by speakers from diverse backgrounds, age groups, gender, and speaking accents. The corpuses contain speech phrases for isolated words, short phrases, paragraphs contents, and long speeches. Most of speech recordings are used from open speech repository that is based on American English accent, available at CMU ARCTIC (Festvox, arctic database). An open source speech dataset consisting contents of children stories (Online audio stories) is also used that consists of 65 children stories of various lengths and recorded by diverse speakers in terms of their age, accent, and gender.

To conduct a case study, we have recorded a speech dataset by 30 speakers (17 male, 13 female) that consists of connected words in the form of digits (5 recordings for each digit by each speaker), short phrases of up to 10 seconds (5 sentences by each speaker) and long phrases of up to 20 seconds (5 paragraph bay each speaker). Although, the proposed TWCST approach is purely based on acoustic features without the transcribed data, yet to prove the concept of language independence, the dataset is recorded for multiple languages that include English, Arabic, and Urdu. For the long speech phrase tracking experiments; a speech corpus from American Rhetoric's (top 100 speeches) (Michael 2001) is used. It is based on hours of speeches recorded by different people on different topics. Moreover, two speech corpuses, Mobio (McCool et al. 2012) and Wolf (Hung and Chittranjan 2010) are requested from IDIAP research institute. These dataset consist hours of speech recordings from variety of speakers in the form of single, binary and group discussions. The details of all datasets are provided in table below.

Table 4-3: Speech Corpuses Used For Performance Evaluation

Corpus name	No. of Speakers	Gender	Length	Availability
Mobio	152	M, F	135 GB	Licence agreement
Wolf	12	M, F	100 GB, 81 hours	Licence agreement
CMU ARCTIC	4	M, F	1150 utterances	Open source
Online Children Stories	65	F	65 stories & poems	Open source
American Rhetoric's	100	M, F	100 speeches	Open source
Case Study Dataset	30	M, F	Connected Digits, Short phrases, Paragraphs	Personal recordings

4.5. Results and Performance Evaluation

Detailed experiments are conducted using aforementioned metrics to validate the performance results. Because of the template frames overlapping, a tolerance of half frame size for the matching decision is set throughout the experiment conduction. In addition to traditional test validation methods, a number of important metrics are added that have been mostly used in the area of binary classification. Figure 4-11 demonstrates an accumulative statistical results comparison of the proposed TWCST approach using different similarity measurement methodologies. Individual performances of adaptive speech tracking based on mean MFCC (KF+ED), DTW with MFCC (KF+DTW), and constrained DTW with MFCC (KF+DTWC) are compared in terms of sensitivity, specificity, matching accuracy, likelihood ratios, F-score, and tracking accuracy. By using the aforementioned simulation setup, 99% sensitivity for the proposed approach is achieved as compared to 45% for DTW. The robustness in true positive detection rates imparts 100% tracking accuracy as desired.

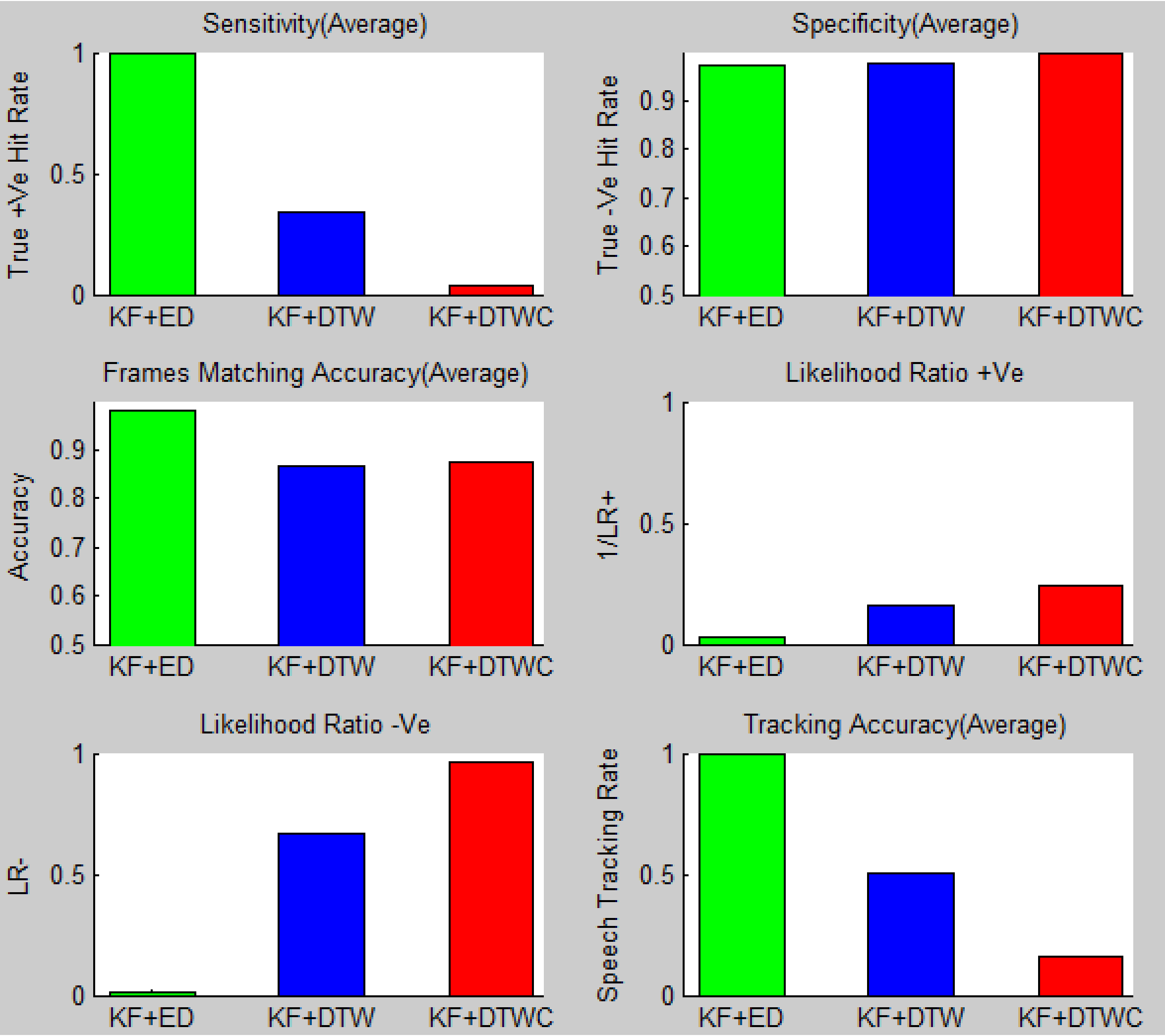


Figure 4-11: Results Comparison of Different Approaches for TWCST

Despite of the fact that the sensitivity and hence the tracking accuracy for DTW and DTWC is very low, yet the perfection in specificity indicates that their performance can be improved using relative threshold. The likelihood ratios (LR+, LR-) are considered as one of the best metrics to measure the diagnostic accuracy. In terms of test and template frames matching, LR presents the probability of a test with test frame match divided by the probability of the same test with test frame mismatch. Larger LR+ consist more information than smaller LR+. On the other hand, smaller LR- consist more information than larger LR-. To simplify the LR values, a relative magnitude is considered by taking the reciprocal of LR+. It can be analysed from Figure 4-11 that the LR- for KF+ED is negligible (0.008) as compared to 0.67 for DTW and 0.96 for DTWC which indicates the robustness of the proposed TWCST. Similarly, F-score is a measure that considers both *precision* and *recall* to measure the system performance. Figure 4-11 demonstrates that the F-score of the proposed TWCST approach is double to DTW based TWCST.

Threshold Setting

It can be observed in Figure 4-11 that the true rejection rates for DTW and DTWC are approximately 100%. The trade-off between true positive hits and true negative rejection rate is based on the threshold value that is used as a decision boundary for test and template frames match/mismatch. Figure 4-12 presents a three dimensional relation among the true positive rate, false positive rate, and threshold values. To set a threshold value for match/mismatch decision boundary, ROC curve is achieved by varying threshold from 0 to 1 with a lag of 0.01. It means that the template frame will be rejected if its matching score with the corresponding test frame is less than the threshold value. The best threshold value (0.85) for the proposed TWCST approach is selected based on a best compromise between sensitivity and specificity as shown in figure below.

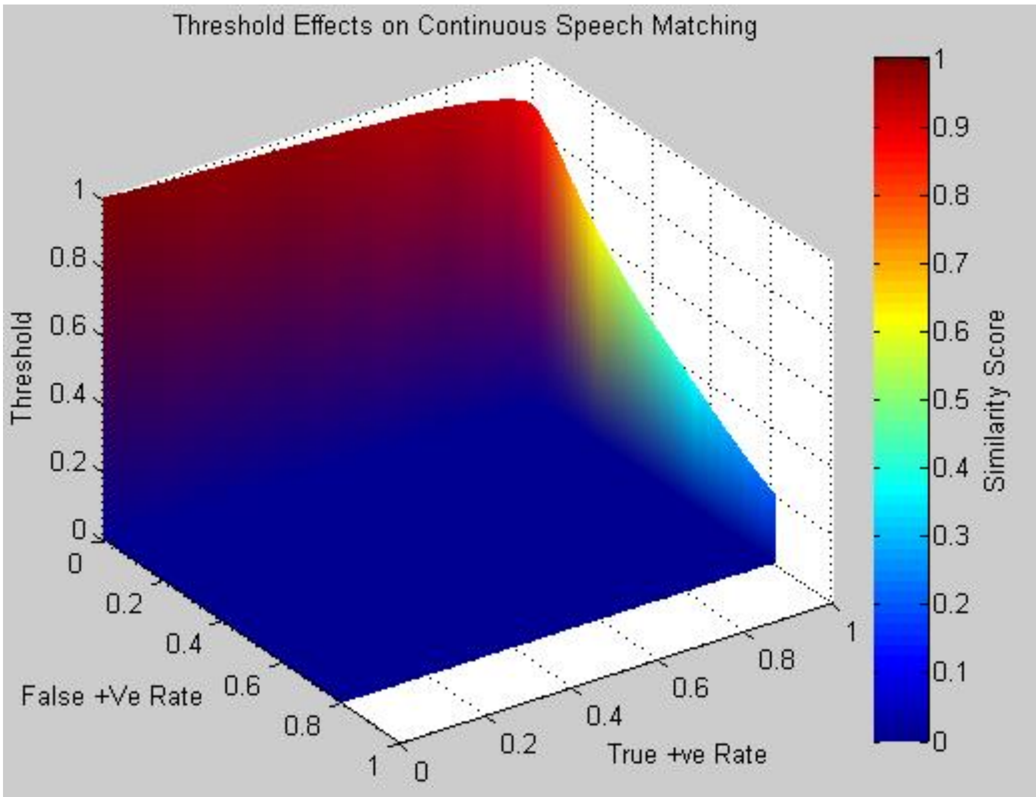


Figure 4-12: Measuring the Best Threshold Value for Similarity Match Decision

Selection of the best threshold values for different TWCST approaches is validated through ROC curve analysis. The area under the curve (AUC) is known as a standard method for ROC curve analysis (Hand and Till 2001). Figure 4-13 demonstrates the ROC analysis for ED using mean MFCC and an enhanced version of DTW (i.e. Segmented DTW) with MFCC. It is clear that the AUC for mean MFCC is greater than DTW based approach which indicates the superiority of the proposed approach over the traditional DTW.

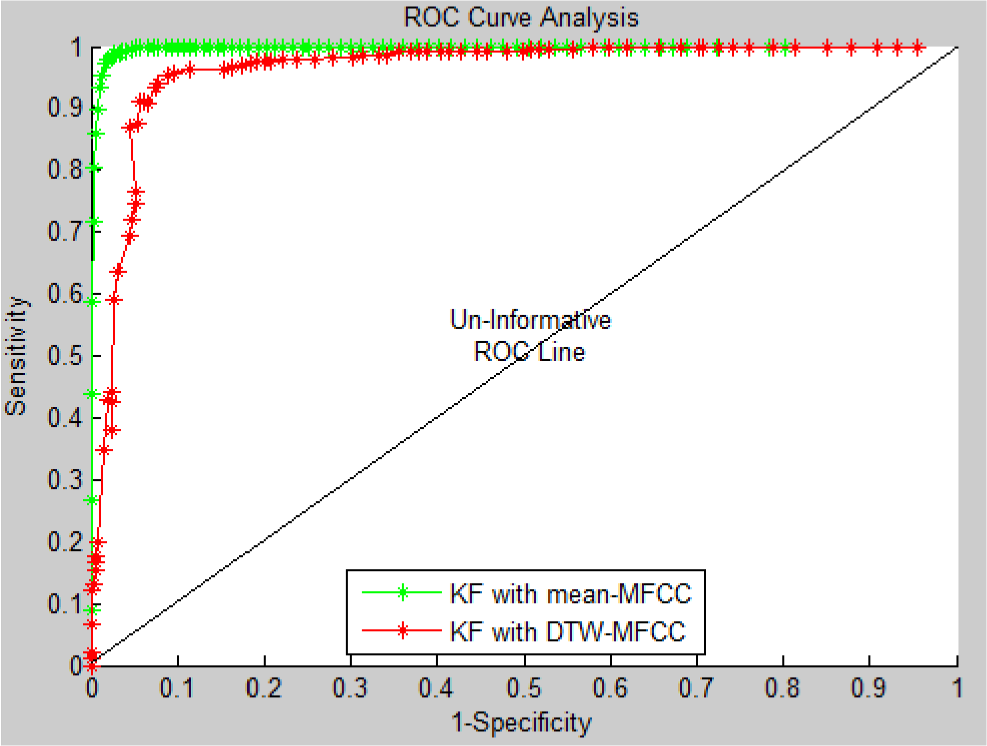


Figure 4-13: Analysis of ROC Curve Based On Varying Threshold Values

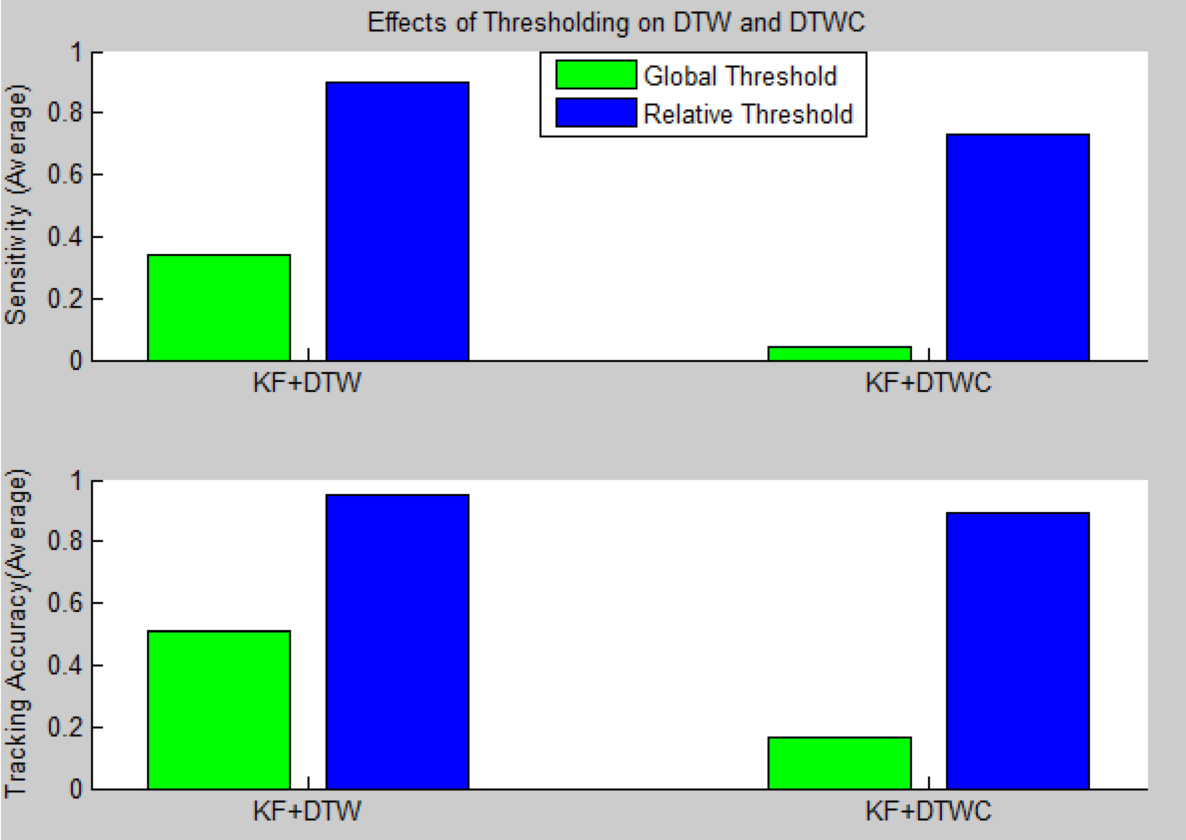


Figure 4-14: Effects of Relative Threshold on the Performance of Different Approaches

Figure 4-14 provides a clear indication of the relative threshold effects for DTW and DTWC based TWCST. It can be observed that the true positive hit rates for DTW and DTWC based approaches are increased from 45% to 89% and 73% respectively. Likewise, the tracking accuracy for both approaches is increased. Despite of the fact that the performance for aforementioned approaches is increased by relative threshold values, yet the statistical results indicate superiority of the proposed adaptive tracking based approach.

Table 4-4 demonstrates the statistical results for TWCST approach using multiple scenarios under diverse circumstances. The performance difference between KF based adaptive framing and search region based non-adaptive approach is presented using the gold standard metrics addressed in Table 4-2. It is observed that the similarity matching and speech tracking performance degrades while using the non-adaptive approach. This is because the proposed adaptive framings approach resolves the issue of dynamic nature of speech in terms of time warping. Also, the use of KF and DSM provides the substitute tracking information that never loses the tracking path when a mismatch or false positive occurs. This proves the reliability of the proposed approach for TWCST. Further detailed statistical results for all aforementioned approaches using the dataset presented in Table 4-3 are addressed in Appendix B and Table 4-4.

Table 4-4: Accumulative Performance Comparison for TWCST

Tracking Approaches		Similarity Measurement Approaches				
		Evaluation Metrics		**Mean-MFCC**	**S-DTW**	**DTWC**
	KF Based Adaptive Tracking (Relative Thresholds)	Sensitivity		0.9918	0.8999	0.7334
		Specificity		0.9726	0.8702	0.9949
		Matching Accuracy		0.9801	0.8958	0.9455
		1/LR+		0.0276	0.1578	0.0071
		LR-		0.0084	0.1229	0.2682
		F-Score		0.9713	0.8487	0.8199
		Tracking Accuracy (%)		1	0.9484	0.8914
		Avg. Execution Time (Sec)		1.3747	3.0419	1.6250
		Type I Error	μ	0.0011	0.0332	8.4183e-05
			σ	0.0016	0.0960	2.1318e-04
		Type II Error	μ	3.1270e-04	0.0282	0.1335
			σ	8.8200e-04	0.0735	0.2487
	SR Based Non- Adaptive Tracking (Static Thresholds)	Sensitivity		0.7299	0.7517	0.1399
		Specificity		0.9856	0.9421	0.9998
		Matching Accuracy		0.9415	0.9121	0.8739
		LR+		0.0182	0.0909	0.0046
		LR-		0.2717	0.2560	0.8603
		F-Score		0.8957	0.8698	0.2587
		Tracking Accuracy (%)		0.6822	0.6991	0.1515
		Type I Error	μ	0.0011	0.0047	1.1904e-06
			σ	0.0016	0.0052	5.2907e-06
		Type II Error	μ	3.1270e-04	0.1462	0.7569
			σ	8.8200e-04	0.2237	0.2035

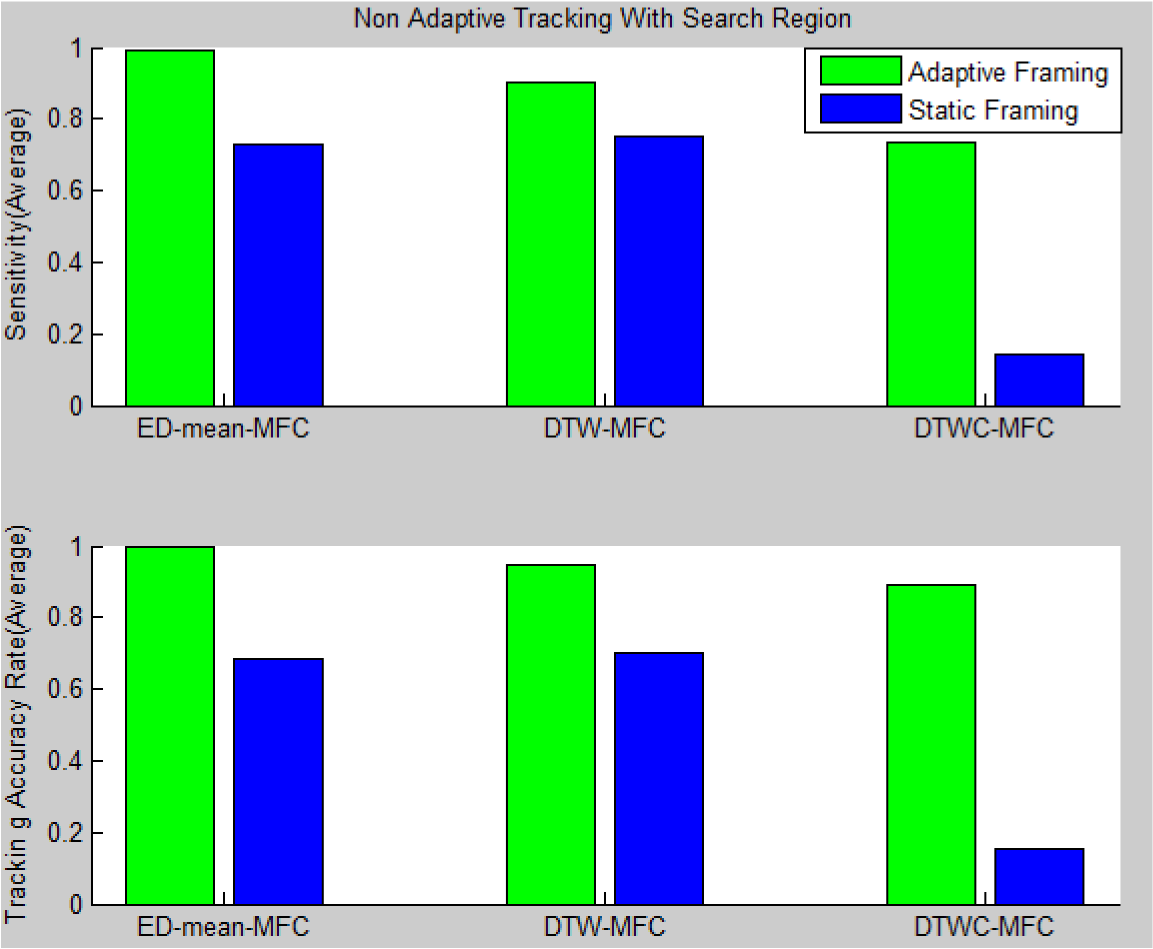


Figure 4-15: Effects of Window Size Adaptation on the Tracking Performance

The figure above explains the degree of diminution in sensitivity and tracking accuracy while excluding the adaptive behaviour from the proposed TWCST approach. A substantial decrement of 27% in the sensitivity and 32% in tracking accuracy can be observed when adaptive framing is changed with non-adaptive static frame size. Likewise, the performances of segmented-DTW and constrained DTW are degraded 15% and 60% respectively in terms of sensitivity. Consequently, the desired speech tracking accuracy is badly affected and is unable to maintain the tracking path information because of mismatches and false negatives. This proves the novelty of the DSM and dynamic frame size concept as introduced in this research study.

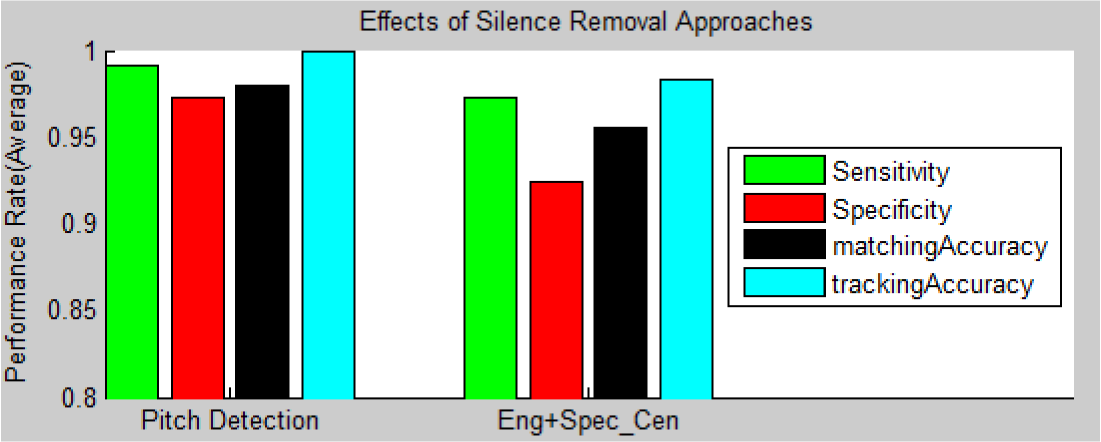


Figure 4-16: Effects of Silence Removal Approaches on the Overall Performance

In addition to DSM and adaptive framing, appropriate selection of a silence removal approach also affects the performance. There are a number of techniques in the literature that uses time and frequency domain features (Energy, zero cross rate, spectral centroid) and pattern recognition methods to remove the silence part of speech utterance (Sahoo and Patra 2014), (Zhang 2014), (Liscombe and Asif 2009), (Saha et al. 2005), (Giannakopoulos 2014), (Sharma and Rajpoot 2013). In the proposed approach for TWCST, an effective method for silence removal is used which is based on robust pitch tracking algorithm (Khan and Holton 2015). In Figure 4-16, the performance for the pitch detection based silence removal is compared with the most commonly used energy and spectral centroid based method. It is clear that the sensitivity, specificity, frames matching accuracy, and speech tracking accuracy are decreased 2%, 5%, 3%, and 2% respectively using traditional silenced removal approach as compared to pitch detection based approach that is introduced very first time in this research study.

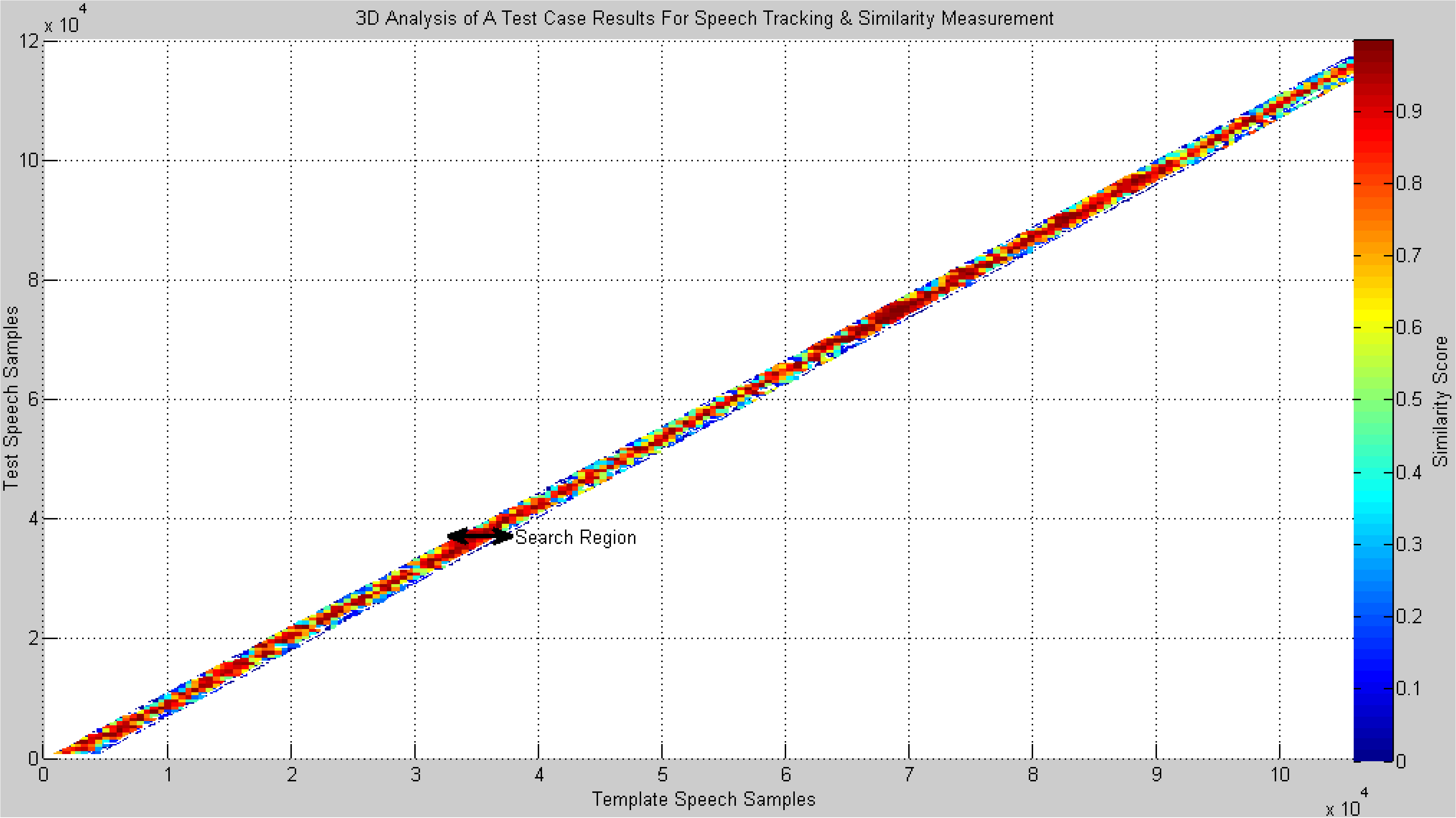


Figure 4-17: 3D Representation of Frame Size and Search Region Adaptation in TWCST for A Test Case

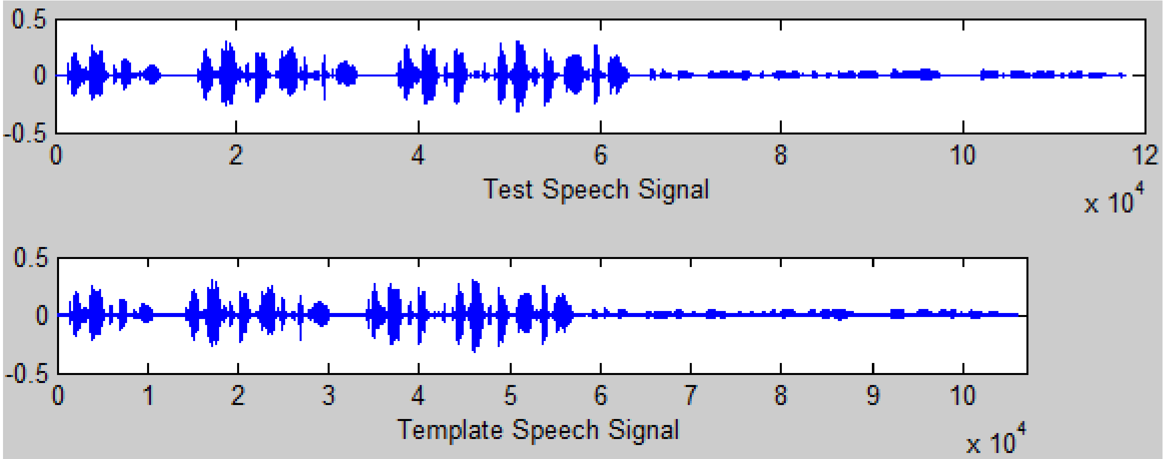


Figure 4-18: A Test Case Speech Data for TWCST

Figure 4-17 demonstrates the proof of concept for TWCST approach using a test case for the speech signals presented in Figure 4-18. The concept of adaptive framing and search region can be observed in 3D representation of the output. Intensity of the colour indicates the match/mismatch score. The dynamics in the frame size can be analysed in the above figure that indicates the adaptive framing in TWCST approach. It can also be observed that there exists at least one frame in the search region that crosses the matching threshold which indicates the robustness in sensitivity leading to consistency in the tracking path. It can also be analysed that the proposed approach resolves the time warping issue in efficient way. For 120,000 samples of test speech as compared to 10,6000 samples for template speech, TWCST approach provides a robust similarity matches between test and template frames without lsoing the tracking path.

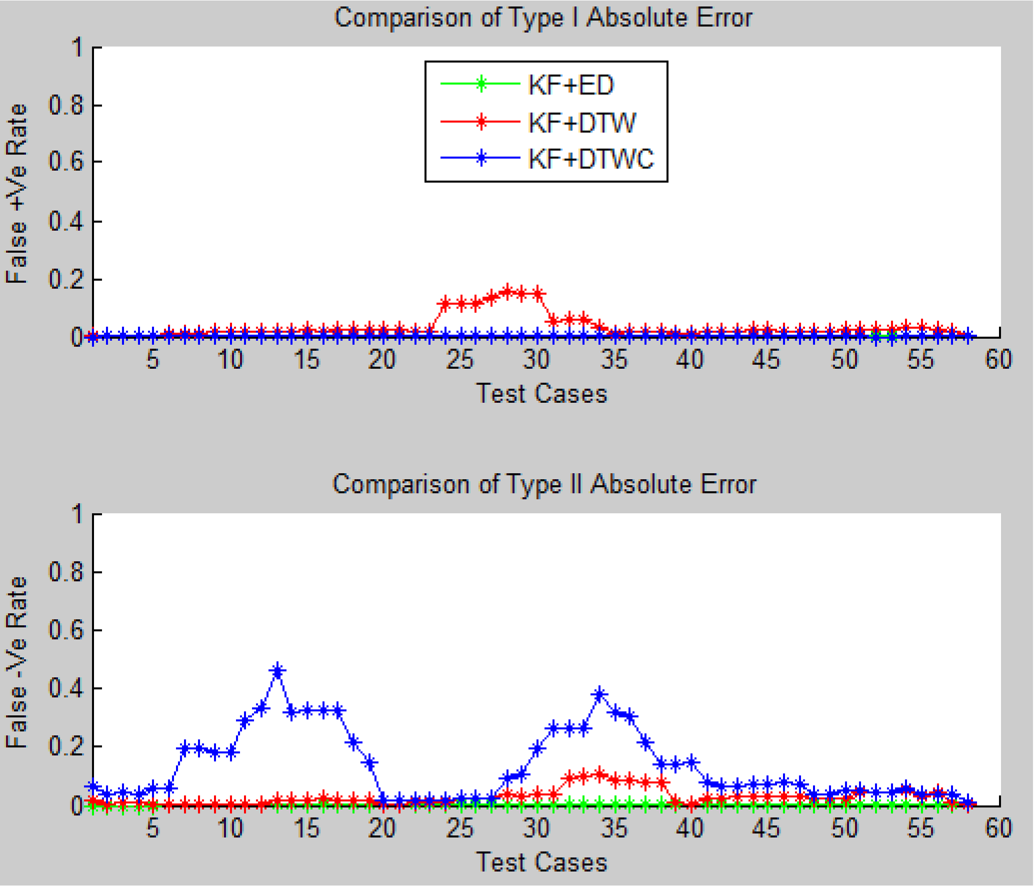


Figure 4-19: Error Rate Analysis for Different Approaches

Absolute Type I and Type II error rates for aforementioned approaches are presented in above figure. It is observed that the both type of errors are approximately zero with the relative thresholds except the constrained DTW. Type I and Type II errors indicates the recogniser failure rates related to FP and FN respectively. These metrics have been represented in a number of ways in the related area including mean square error and absolute erros as most commonly used. Table 4-4 demonstrates ‘µ’ (mean) and ‘σ’ standard deviation for both types of error for different approaches while using the same speech dataset.

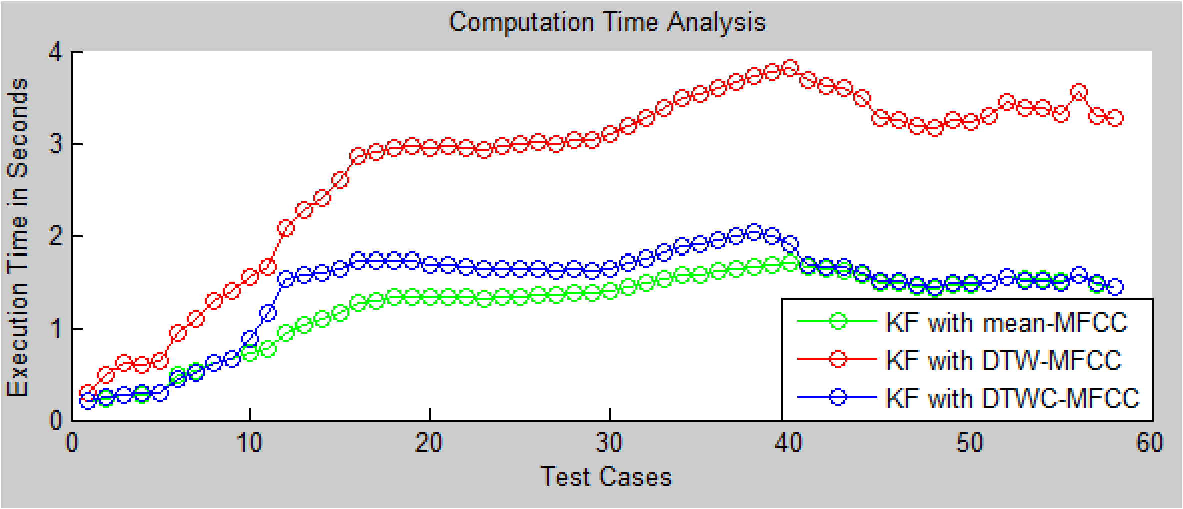
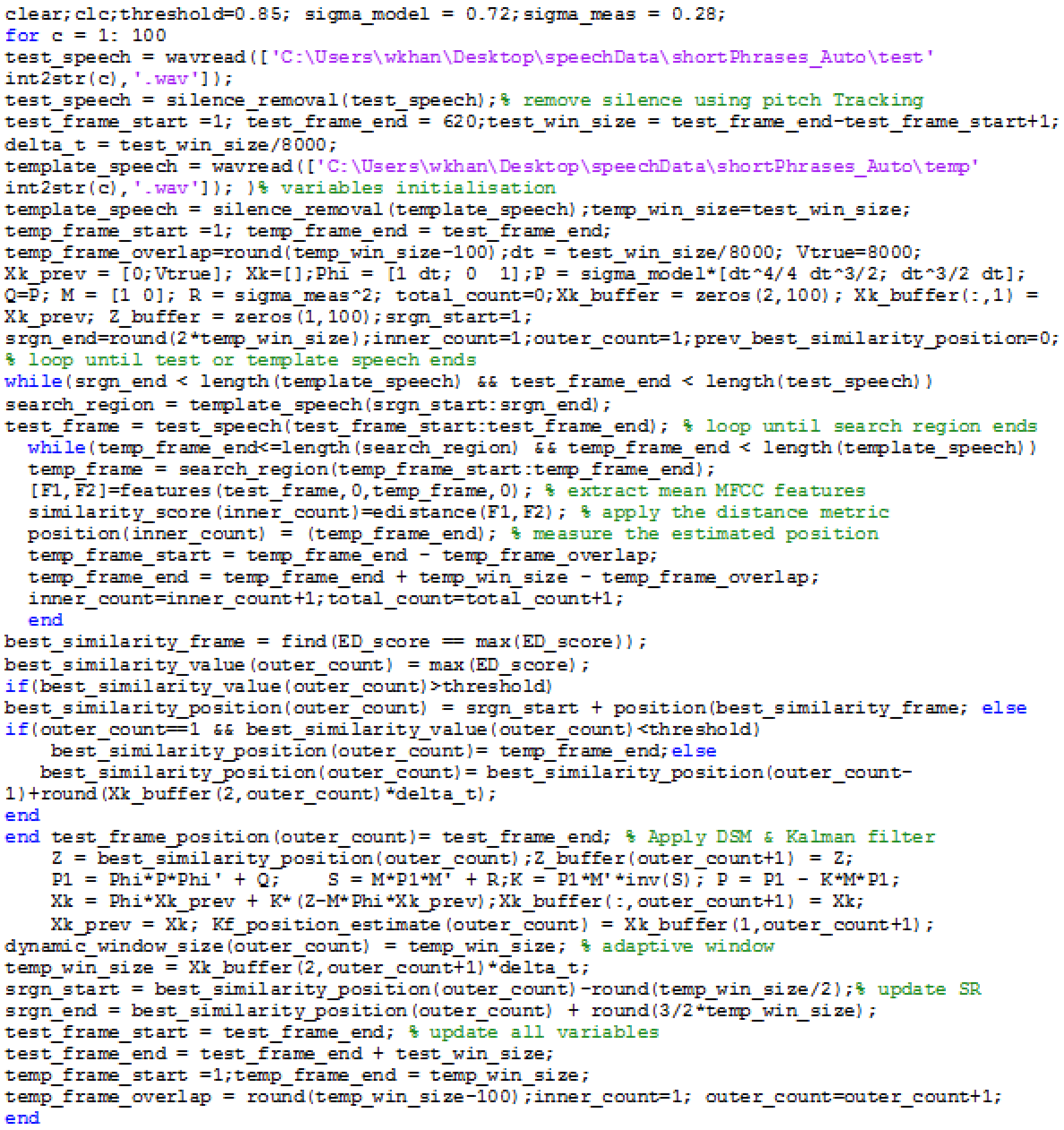


Figure 4-20: Comparison of Computation Costs of Multiple Approaches

Computation time is also an important factor that is needed to be analysed for aforementioned approaches. Table 4-4 shows the average computational costs for these approaches whereas; Figure 4-20 demonstrates variations in computational time for varying lengths of test and template speech signals. It is observed that the KF based approach have a lead over the conventional approaches. This is because of computation time in DTW is linear with the number of frames to be searched through (Cheng-Tao et al. 2014) (detailed description in Chapter 3, Section 3.4). Despite of the fact that constrained DTW reduces the average time complexity to 1.6 seconds, yet its computation time is greater than mean MFCC based approach (1.3 sec) because of two dimensional local search processes (Chapter 3, Section 3.4). Finally, a Matlab script is presented in Figure 4-21 that reflects the processing flow shown in Figure 4-1. Detailed Matlab script for TWCST approach using all aforementioned methodologies is provided in Appendix D.



```
clear;clc;threshold=0.85; sigma_model = 0.72;sigma_meas = 0.28;
for c = 1: 100
test_speech = wavread(['C:\Users\wkhan\Desktop\speechData\shortPhrases_Auto\test'
int2str(c),'.wav']);
test_speech = silence_removal(test_speech);% remove silence using pitch Tracking
test_frame_start =1; test_frame_end = 620;test_win_size = test_frame_end-test_frame_start+1;
delta_t = test_win_size/8000;
template_speech = wavread(['C:\Users\wkhan\Desktop\speechData\shortPhrases_Auto\temp'
int2str(c),'.wav']); )% variables initialisation
template_speech = silence_removal(template_speech);temp_win_size=test_win_size;
temp_frame_start =1; temp_frame_end = test_frame_end;
temp_frame_overlap=round(temp_win_size-100);dt = test_win_size/8000; Vtrue=8000;
Xk_prev = [0;Vtrue]; Xk=[];Phi = [1 dt; 0  1];P = sigma_model*[dt^4/4 dt^3/2; dt^3/2 dt];
Q=P; M = [1 0]; R = sigma_meas^2; total_count=0;Xk_buffer = zeros(2,100); Xk_buffer(:,1) =
Xk_prev; Z_buffer = zeros(1,100);srgn_start=1;
srgn_end=round(2*temp_win_size);inner_count=1;outer_count=1;prev_best_similarity_position=0;
% loop until test or template speech ends
while(srgn_end < length(template_speech) && test_frame_end < length(test_speech))
search_region = template_speech(srgn_start:srgn_end);
test_frame = test_speech(test_frame_start:test_frame_end); % loop until search region ends
  while(temp_frame_end<=length(search_region) && temp_frame_end < length(template_speech))
  temp_frame = search_region(temp_frame_start:temp_frame_end);
  [F1,F2]=features(test_frame,0,temp_frame,0); % extract mean MFCC features
  similarity_score(inner_count)=edistance(F1,F2); % apply the distance metric
  position(inner_count) = (temp_frame_end); % measure the estimated position
  temp_frame_start = temp_frame_end - temp_frame_overlap;
  temp_frame_end = temp_frame_end + temp_win_size - temp_frame_overlap;
  inner_count=inner_count+1;total_count=total_count+1;
  end
best_similarity_frame = find(ED_score == max(ED_score));
best_similarity_value(outer_count) = max(ED_score);
if(best_similarity_value(outer_count)>threshold)
best_similarity_position(outer_count) = srgn_start + position(best_similarity_frame; else
if(outer_count==1 && best_similarity_value(outer_count)<threshold)
    best_similarity_position(outer_count)= temp_frame_end;else
   best_similarity_position(outer_count)= best_similarity_position(outer_count-
1)+round(Xk_buffer(2,outer_count)*delta_t);
end
end test_frame_position(outer_count)= test_frame_end; % Apply DSM & Kalman filter
    Z = best_similarity_position(outer_count);Z_buffer(outer_count+1) = Z;
    P1 = Phi*P*Phi' + Q;     S = M*P1*M' + R;K = P1*M'*inv(S); P = P1 - K*M*P1;
    Xk = Phi*Xk_prev + K*(Z-M*Phi*Xk_prev);Xk_buffer(:,outer_count+1) = Xk;
    Xk_prev = Xk; Kf_position_estimate(outer_count) = Xk_buffer(1,outer_count+1);
dynamic_window_size(outer_count) = temp_win_size; % adaptive window
temp_win_size = Xk_buffer(2,outer_count+1)*delta_t;
srgn_start = best_similarity_position(outer_count)-round(temp_win_size/2);% update SR
srgn_end = best_similarity_position(outer_count) + round(3/2*temp_win_size);
test_frame_start = test_frame_end; % update all variables
test_frame_end = test_frame_end + test_win_size;
temp_frame_start =1;temp_frame_end = temp_win_size;
temp_frame_overlap = round(temp_win_size-100);inner_count=1; outer_count=outer_count+1;
end
```

Figure 4-21: Matlab Script for TWCST

4.6. Summary

In this chapter, a comprehensive overview of the research contribution towards TWCST approach and its comparison with existing approaches is presented. For the first time, the concept of dynamic frame size is introduced using a dynamic state model that is based on object’s linear motion. A detailed flowchart for sequential processing of TWCST is presented along-with the mathematical formulation of a

DSM and KF in the proposed approach. Kalman filter that has successfully been used for object localisation is deployed as a recursive feedback system that fuses the location information from more than one resource. The deployment of KF and DSM is novel in speech tracking. It empowers not only the decision of match/mismatch but also provides a backup to recover the tracking path in case of mismatch or false negatives. In addition, a novel approach for removing the silence segments from the input speech signal is introduced that is based on pitch tracking in speech signal. Multiple speech corpuses are used for evaluating the proposed approach. Performance of the newly introduced TWCST approach is evaluated using the gold standards metrics for binary classifier (i.e. match/mismatch) as well as for speech tracking. In addition to speech tracking performance, the new approach is compared with the existing works in terms of similarity matching and computational cost analysis. Finally, detailed results are presented in form of statistics and graphical representations.

5. SPEECH TRACKING USING DYNAMIC NOISE FILTRATION BY WAVELET DECOMPOSITION

5.1. Scope

This chapter presents an alternative approach for TWCST that uses the Wavelet Decomposition (WD) for time frequency representation of speech signal. In Section 5.2, the process of WD is presented followed by the introduction of a dynamic noise filtration method to enhance a speech signal by filtering out the unnecessary information from speech signal. Section 5.3 addresses the performance comparison for different scenarios including silence removal techniques, static frame size, and adaptive framing based TWCST approach. Section 5.4 demonstrates a detailed discussion on the advantages of proposed TWCST approaches over the existing speech signal matching techniques while considering their practical and theoretical aspects. Finally, a brief summary of the chapter is presented.

5.2. Dynamic Noise Filter and Time Warped Continuous Speech Tracking

Research work presented in (Khan et al. 2014) introduces an alternative approach for the TWCST that is addressed in Chapter 4. The major differences for alternative TWCST approach include:

- WD for spectral analysis
- Dynamic noise filtration
- Wavelets energy as features vector

The WD is superior over the FFT in terms of defining a particularly useful class of time-frequency distributions which specify complex amplitude versus time and frequency for any signal (Fugal 2010). Wavelets express signals as sums of wavelets and their dilations and translations. They act in a similar way as FFT but can approximate signals which contain both; large and small frequencies as well as discontinuities. This is due to the fact that wavelets do not use a fix time frequency window as in case of FFT or STFT.

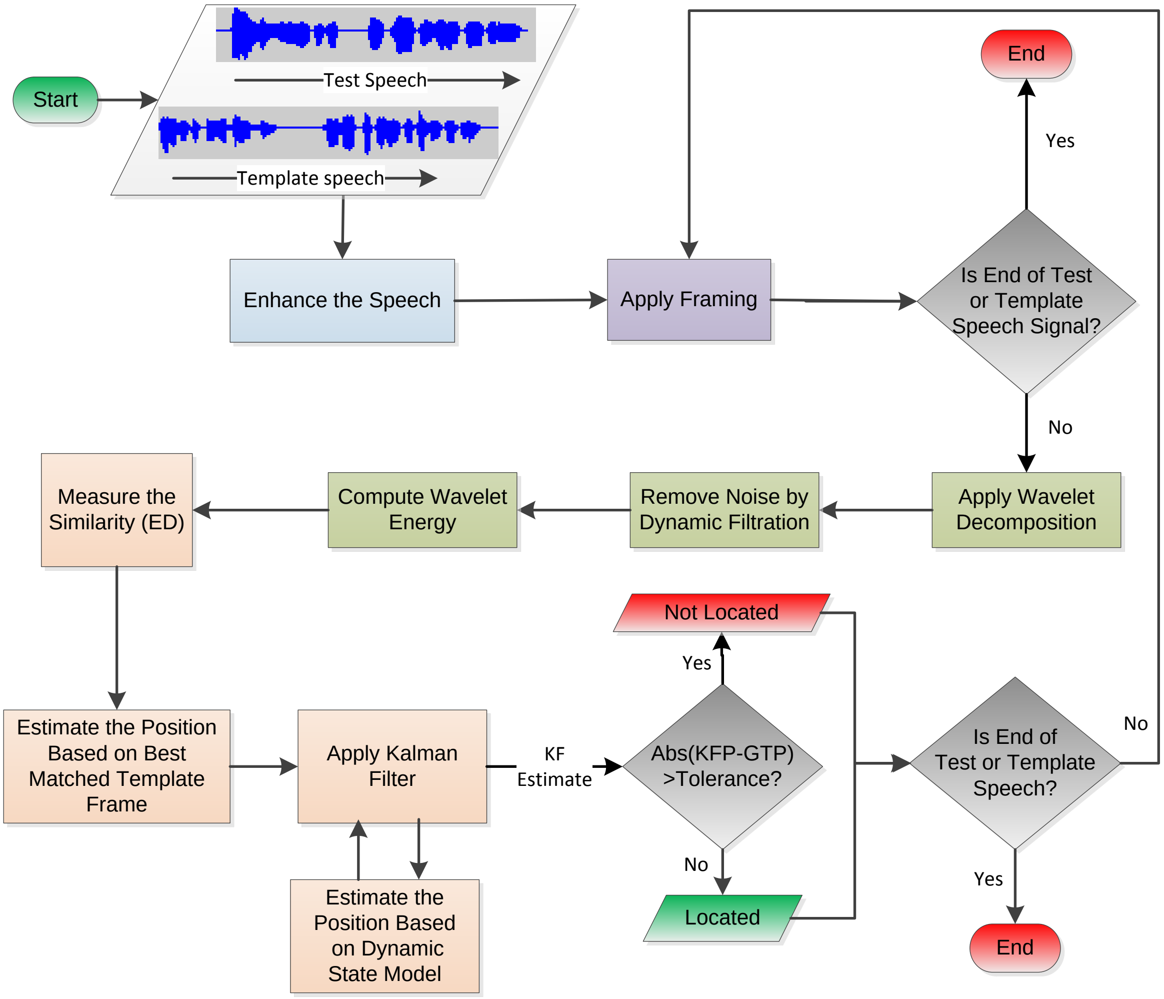

Start
Test Speech
Template speech
End
Yes
Enhance the Speech
Apply Framing
Is End of Test or Template Speech Signal?
No
Measure the Similarity (ED)
Compute Wavelet Energy
Remove Noise by Dynamic Filtration
Apply Wavelet Decomposition
Not Located
Yes
Estimate the Position Based on Best Matched Template Frame
Apply Kalman Filter
KF Estimate
Abs(KFP-GTP) >Tolerance?
Is End of Test or Template Speech?
No
No
Located
Yes
End
Estimate the Position Based on Dynamic State Model

Figure 5-1: TWCST Approach Using Wavelets Based Dynamic Filter

The underlying principle of wavelets is to analyse according to varying scale. Hence, overall phenomenon of WD provides an advantage over FFT that has the issue of poor time-frequency resolution. That is, if window size is kept small, the frequency resolution will be poor and vice versa. Figure 5-1 represents the sequential flow of various processes that are amalgamated to produce a new

TWCST approach. It can be analyse in Figure 5-1 that most of its processes are identical to the previous approach (Figure 4-1) except the WD, dynamic noise filter, and feature extraction. Therefore, only those aspects are discussed in this chapter that are different from previous approach followed by the detailed evaluation of their impacts on the statistical results.

5.2.1. Wavelet Decomposition

Discrete time signal decomposition was introduced by Croiser, Esteban, and Galand in 1976. Later on, a number of variations have been introduced to decompose the digital signal with different approaches and terminologies. However, the main idea is same as CWT that is already discussed in Chapter 2, Section 2.2.2. As compared to CWT that is computed by scale changing, shifting the window in time, multiplying the signal and integrating over all time, WD uses filters for different cut-off frequencies to analyse the signal at different scales (Policar 2006). In the first step, test and template speech signals are passed through a series of high pass and low pass filters to analyse the high and low frequencies respectively.

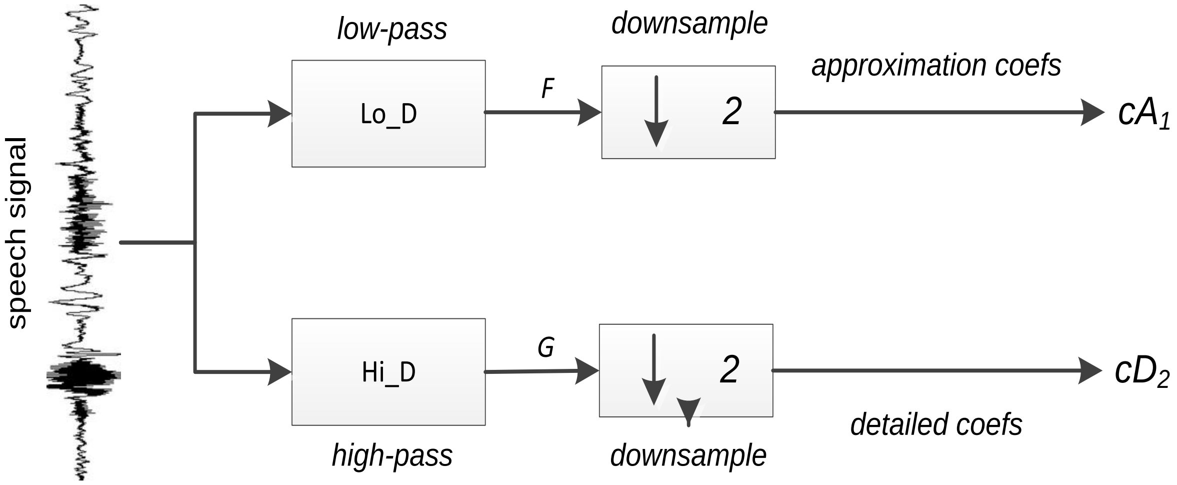


Figure 5-2: Wavelet Decomposition

Where *'F'* and *'G'* are the low-pass and high-pass filtered speech signals respectively, cA_1 are the approximation coefficients and cD_1 are the detailed coefficients at level 1. This procedure runs recursively and the approximation coefficients are further decomposed until the desired number of levels. Thus, for next step, cA_1 will be decomposed into cA_2 and cD_2 and so on as shown in figure below.

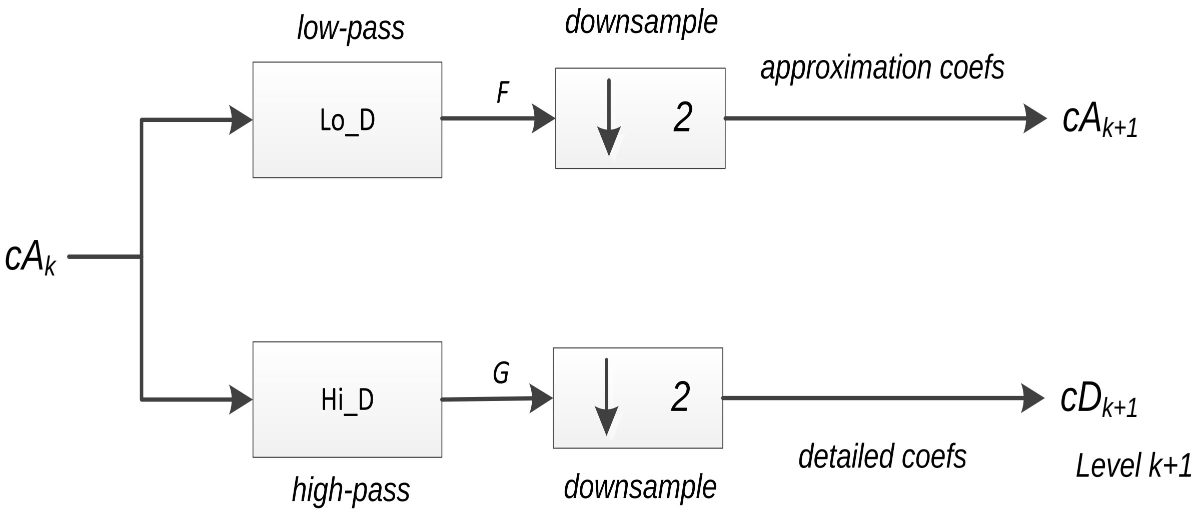


Figure 5-3: Decomposition of the Approximation Coefficients

The low-pass and high-pass filters use the mathematical operation of convolution of the speech signal with the impulse response of filter and can be expressed as:

$$(x*h)[n] = \sum_{k=-\infty}^{\infty} x[k].\, h[n-k] \quad \text{5-1}$$

Where 'x' is the speech signal, 'h' is the filter impulse response, 'k' represents the sample number, and 'n' is the index. The frequencies above the half of the highest frequency in the speech signal are removed by half band low-pass filter. Scale of the signal is doubled by subsampling the signal by two. This procedure can mathematically be represented as:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k].\, h[2n-k] \quad \text{5-2}$$

In WD, speech signal is analysed at different frequency bands by decomposing the signal into the approximation and detailed coefficients using a scaling and wavelet processes. Scaling function produces the task of low-pass filter while wavelet function performs the high-pass filtration on the input speech signal. A time domain speech signal is thus decomposed into different levels (i.e. frequency bands) by recursive low-pass and high-pass filtration processes. Let's $'g'$and $'h'$represent the high-pass and low-pass filters respectively for input speech signal$'x'$, then a single level WD expressed in Equation 5-2 can be rewritten as:

$$y_{high}\left[n\right]=\sum_{k=-\infty}^{\infty}x[k].\,g\left[2n-k\right] \qquad 5\text{-}3$$

$$y_{low}\left[n\right]=\sum_{k=-\infty}^{\infty}x[k].\,h\left[2n-k\right] \qquad 5\text{-}4$$

Where y_{high} and y_{low} are the high-pass and low-pass filters output respectively after subsampling by a factor of 2 for each index 'n'. The decomposition process reduces the time resolution by half because of subsampling while increasing the frequency resolution to double. This procedure is also known as sub-band coding which runs recursively for further decomposition till desired level as shown in Figure 5-2 and 5-3.

5.2.2. Dynamic Noise Filtration

In the second step, a dynamic noise filter is applied to time and frequency domain speech signal. Wavelet based signal decomposition has an advantage over the other spectral analysis techniques because of its multi-scale representation of function. A WD analyses the function at various levels of resolution and provides a simultaneous time-frequency representation of input speech signal. This representation empowers the wavelet's superiority because of its efficiency for localising the frequency in time domain along with the correlation matrix as a third dimension.

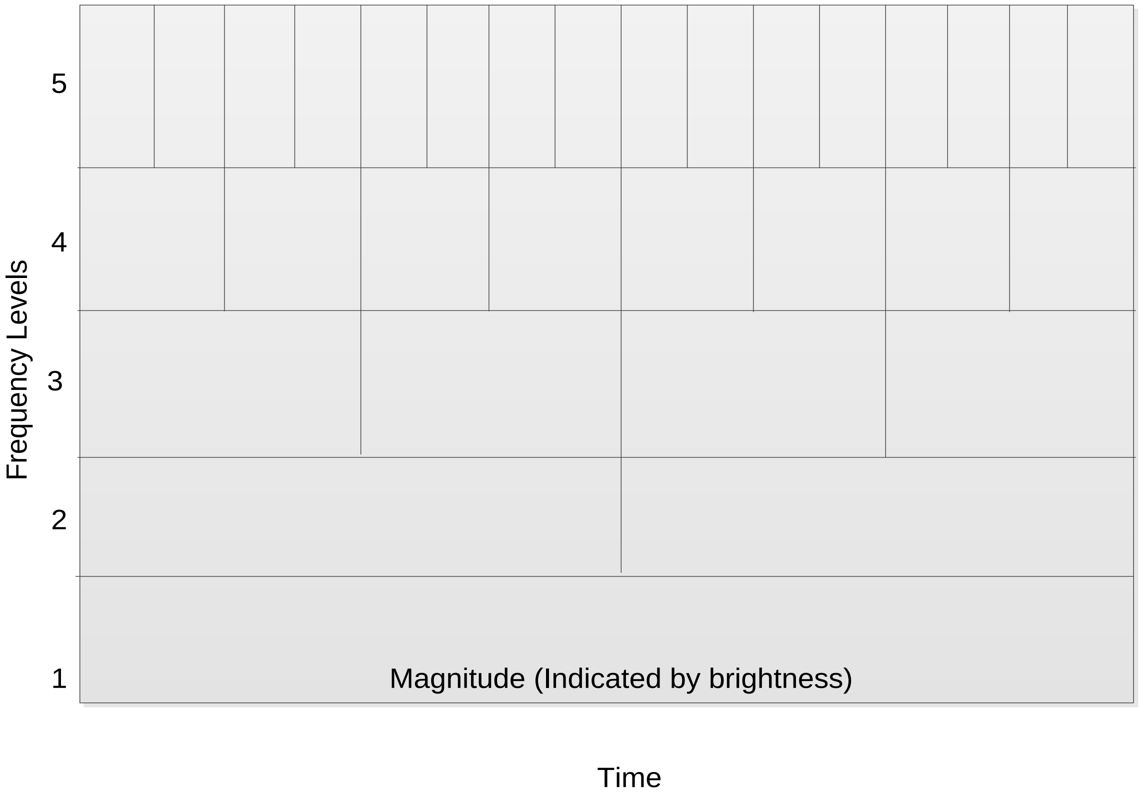


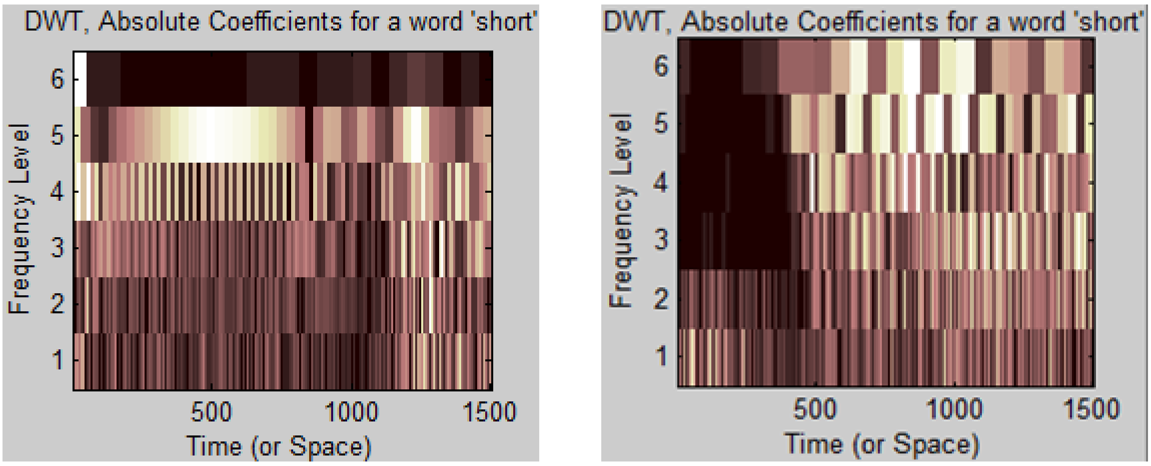
Figure 5-4: A 3-Dimensional Representation of WD Output

A three dimensional representation of WD output is demonstrated in Figure 5-4. It can be observed in above figure that the WD output can be analysed simultaneously in time-frequency domain. Thus the task of noise filtration can easily be performed by applying a threshold constraint on the energy magnitude. Approximation and detailed coefficients possessing the magnitude values below a predefined threshold are filtered out while remaining coefficients are integrated together and forwarded for further processing. After the noise filtration, energy in a frequency level is measured by integrating the intensity magnitudes over time and can be represented as:

$$E_{Scale} = \frac{\sum_{i=1}^{n} \left(y_i^{Scale} \right)^2}{n} \qquad 5\text{-}5$$

Where 'n' represents the total number of coefficients in current frequency scale, E_{Scale} is the total energy measure for specific scale, and 'y' represents the output coefficients produced by WD for the current scale. The above procedure is applied to each scale and corresponding energy vector is measured. For the TWCST approach, speech signal is decomposed up to seven scales. This is because the sampling frequency of input speech is set to 8000 Hz and following the Nyquist sampling, 4000 Hz would be enough to analyse the speech signal. This implies that the signal is analysed between 64 Hz to 4000 Hz which covers the lower and upper bounds of human voice frequency ranges. After calculating the normalised energy for each frequency band, a threshold value (0.5) is applied to each frequency band to filter out the unnecessary scales.

A WD approach has been successfully used as a spectral analysis tool (Policar 2001). It can effectively compress the information about the non-stationary into a piece of local information. Moreover, it reveals the scale-wise organization of singularities, thus allowing for the selection of the interesting strongest events (Young 2008). As speech signal is dynamic in nature, for multiple instances of a same speech utterance for same speaker, different frequency magnitude values may be produced by WD.



(a) (b)

Figure 5-5: Dynamic Level Time-Frequency Domain Noise Filtration (Khan et al. 2014)

Figure 5-5 illustrates the three dimensional representation and noise filtration example for a test case. The WD is performed on an isolated speech utterance. It can be analysed that in Figure 5-5(a), most of the frequencies (i.e. colour intensity indicates the magnitude values for a specific frequency band) are available in level 3, 4, and 5. Therefore, level 1, 2, and 6 can be filter out as there is no clue of useful information that can be used for feature extraction. However; it can be observed in Figure 5-5(b) that for same speech utterance by same speaker but recorded at different time, level 6 can't be filtered out because of the high energy components existence in this level. In addition to frequency domain noise filtration, same process is repeated for time domain components as well. For example, initial 500 coefficients can be filter out in Figure 5-5(b) as they possess extremely low energy magnitude.

After the dynamic noise filtration and energy calculation for test and template speech frames, a similarity score is calculated using the Euclidean distance. Details for calculation of similarity score is already presented in Chapter 4, Section 4.3.3. Likewise the previous approach, a DSM is formulated and the KF is used as the recursive feedback system which is described in Chapter 4, Section 4.3.4. Procedure of KF tuning, frame size adaptation, and search region is similar to previous approach for TWCST (Section 4.3).

5.3. Performance Evaluation

Experiments are conducted using the statistical metrics described in Chapter 4, section 4.4 to validate the performance results. The experimental setup is already discussed in Chapter 4, section 4.4 along with the simulation tools and speech corpuses (Table 4-3). Likewise the previous approach, half of the template frame size is set as a tolerance values for tracking accuracy measurement throughout the experimental analysis. Figure 5-6 demonstrates an accumulative statistical results comparison of the proposed TWCST approach using KF based adaptive frame size and search region based static framing. Performances for both scenarios are compared in terms of sensitivity, specificity, matching accuracy, likelihood ratios, F-score, and tracking accuracy. The likelihood ratios (i.e. LR+, LR-) are considered one of the best metrics to measure the diagnostic accuracy. In terms of test and template frames matching, LR presents the probability of a test with test frame match divided by the probability of the same test with test frame mismatch. Larger LR+ consist more information than smaller LR+. On the other hand, smaller LR- consist more information than larger LR-. To simplify the LR values, a relative magnitude is considered by taking the reciprocal of LR+. Similarly, F-score is a measure that considers both *precision* and *recall* to measure the system performance.

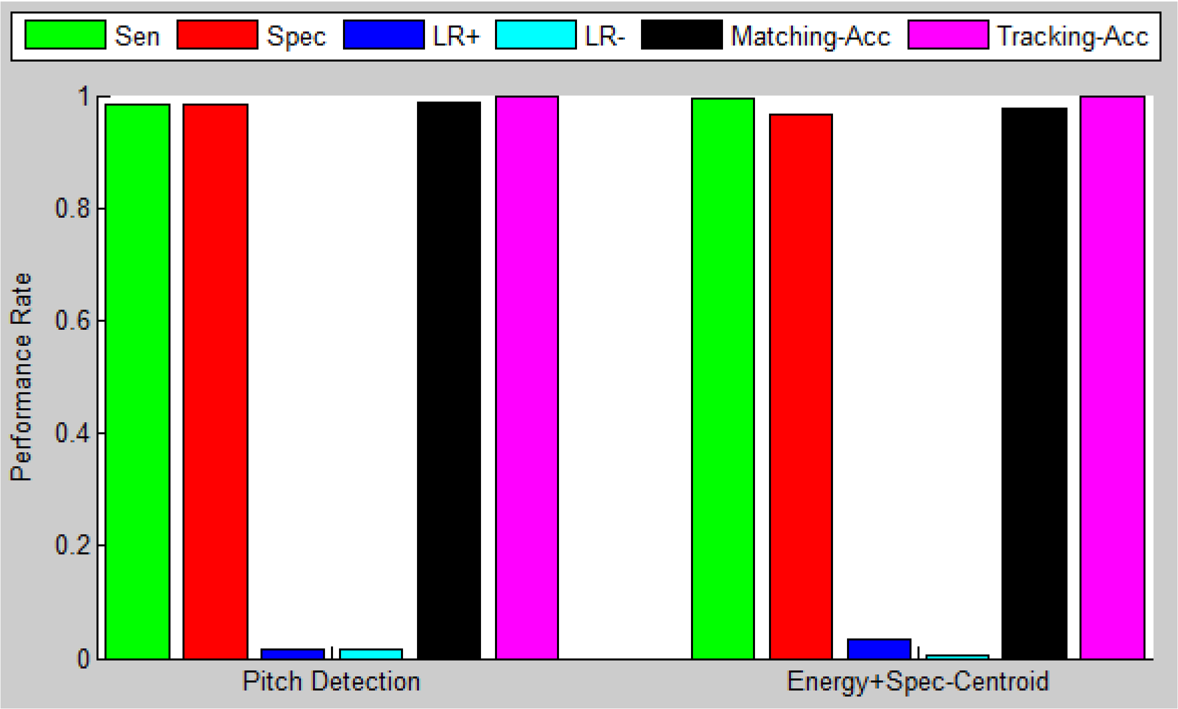


Figure 5-6: Statistical Results Comparison for TWCST Approaches

It can be observed that the statistical results for pitch detection based silence removal are almost identical with the traditional silence removal approach. This proves the robustness of the dynamic filter which discards unnecessary coefficients from the speech signal. Hence, the poor performance of a silence removal approach does not affect the tracking accuracy.

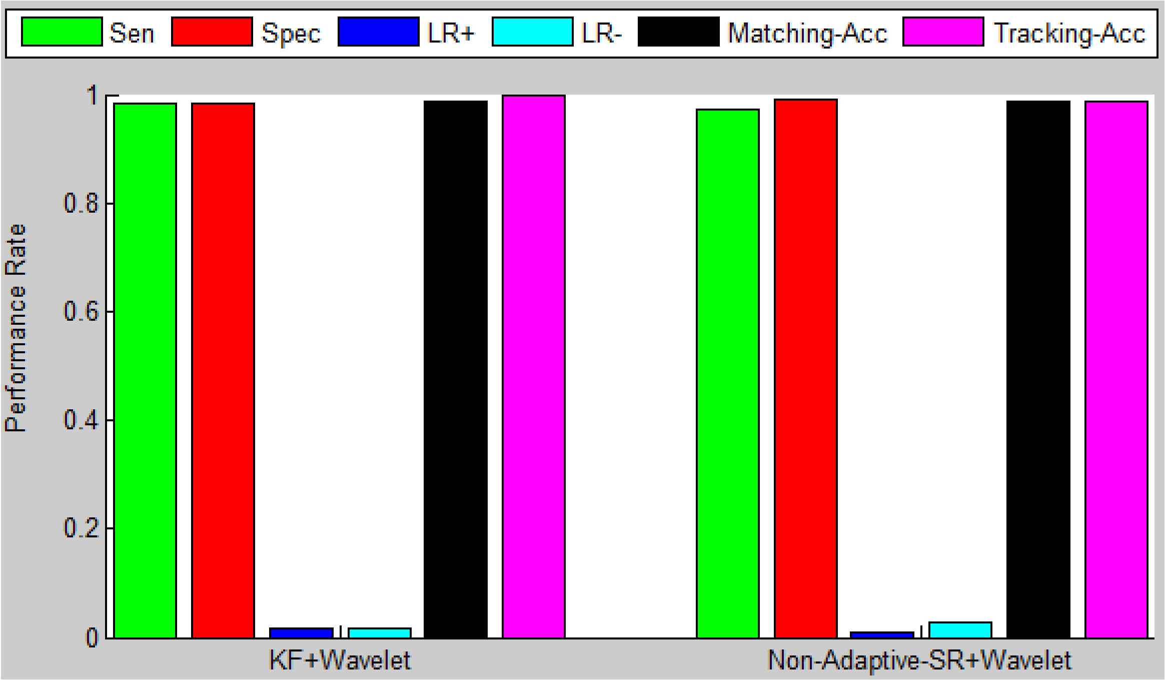


Figure 5-7: Statistical Results Comparison for Static and Adaptive Framing

Figure 5-7 shows the statistical results for static and adaptive framing based TWCST approaches. It is observed that the statistical metrics indicate almost equal performance for both approaches except the sensitivity which is slightly (2%) decreased for static framing. As the LR- and tracking accuracy are directly related to the sensitivity, there is a minor decrement in the performance for these metrics. The robustness in true positive detection rates imparts 100% tracking accuracy as desired. The trade-off between true positive hits and true negative rejection rate is based on the threshold value that is use as a decision boundary for test and template frames match/mismatch. Figure 5-8 presents the relation between true positive rate, false positive rate, and threshold values. To set a threshold value for match/mismatch decision boundary, a ROC curve is achieved by varying threshold from 0 to 1 with a lag of 0.01. It means that the template frame will be rejected if its matching score with the corresponding test frame is less than the threshold. The best threshold value (0.75) for the proposed wavelet based TWCST approach is selected by a compromise between sensitivity and specificity.

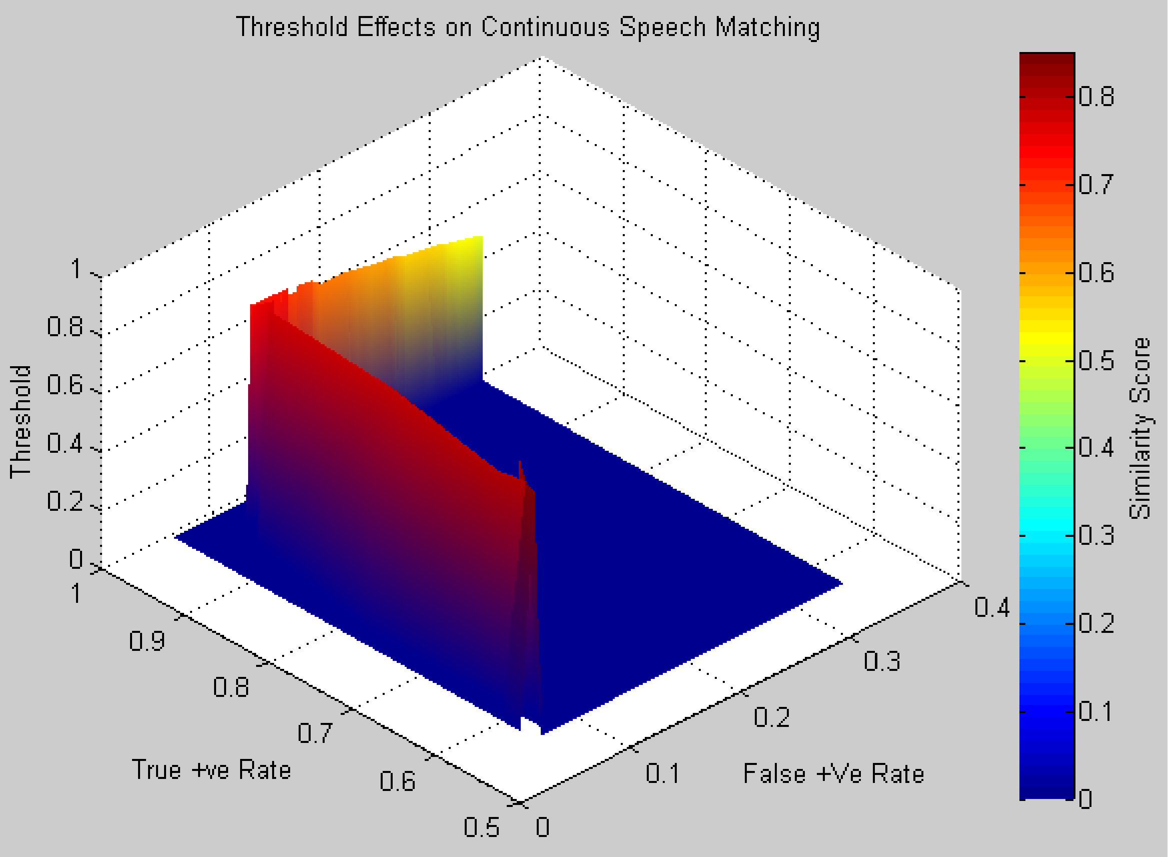


Figure 5-8: Measuring the Best Threshold Value for Similarity Match Decision

Absolute Type I and Type II error rates for aforementioned scenarios are presented in Figure 5-9. Type I and Type II errors indicates the recogniser failure rates related to FP and FN respectively. These metrics have been represented in a number of ways in the related area including mean square error and absolute erros as most commonly used. It is observed that the Type I error is approximately zero in all scenarios which shows the robustness of true negative rate. However, Type II error increased slightly (from 0.01 to 0.02) for static frame based CTWST approach that proves the superiority of KF based frame size adaptation for tracking accuracy.

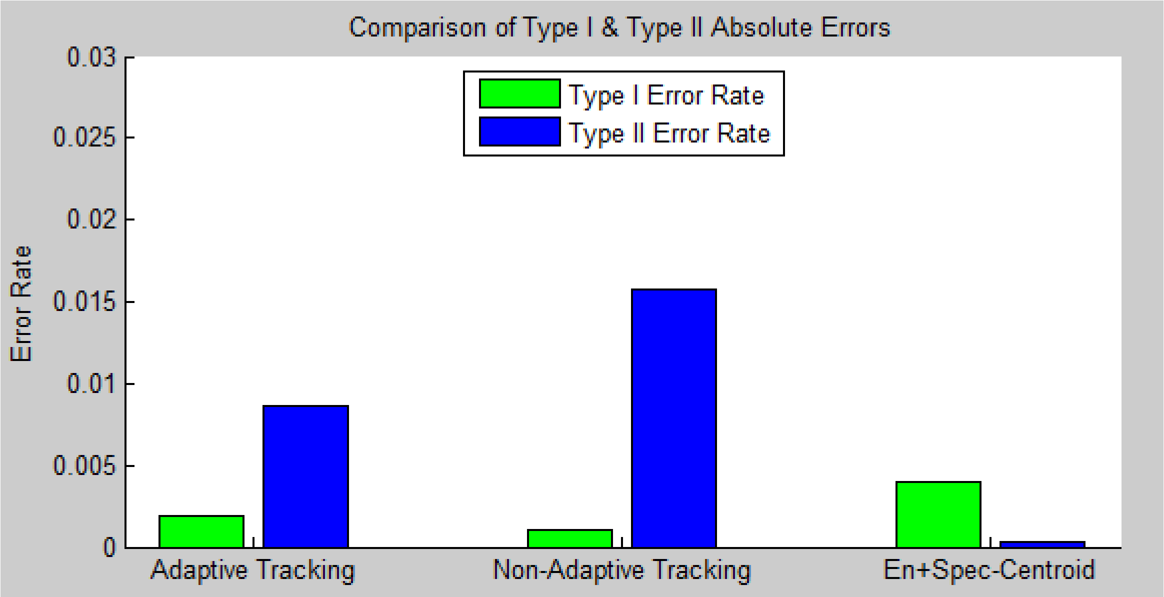


Figure 5-9: Error rate analysis for different approaches

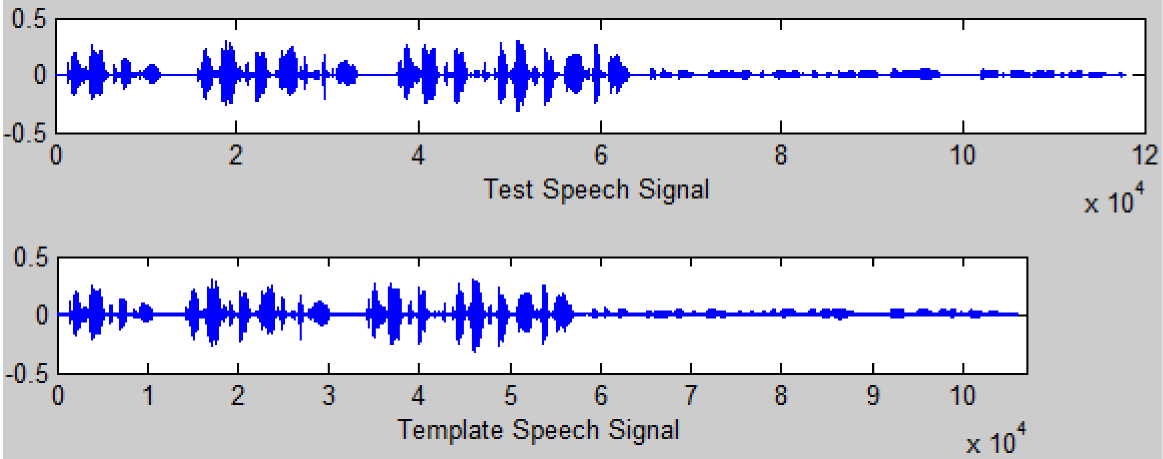


Figure 5-10: A Test Case Speech Data for Wavelet Based TWCST Approach

Figure 5-11 demonstrates the proof of concept by a test case for TWCST using the speech signals shown in Figure 5-10. The concept of adaptive framing and search region can be observed in 3D representation of the output. The colour intensity indicates the match/mismatch score. The dynamics in the frame size can be analysed in Figure 5-11 that indicates the adaptive framing in TWCST approach. It can also be observed that there exists at least one frame in the search region that crosses the matching threshold which indicates the robustness in sensitivity leading

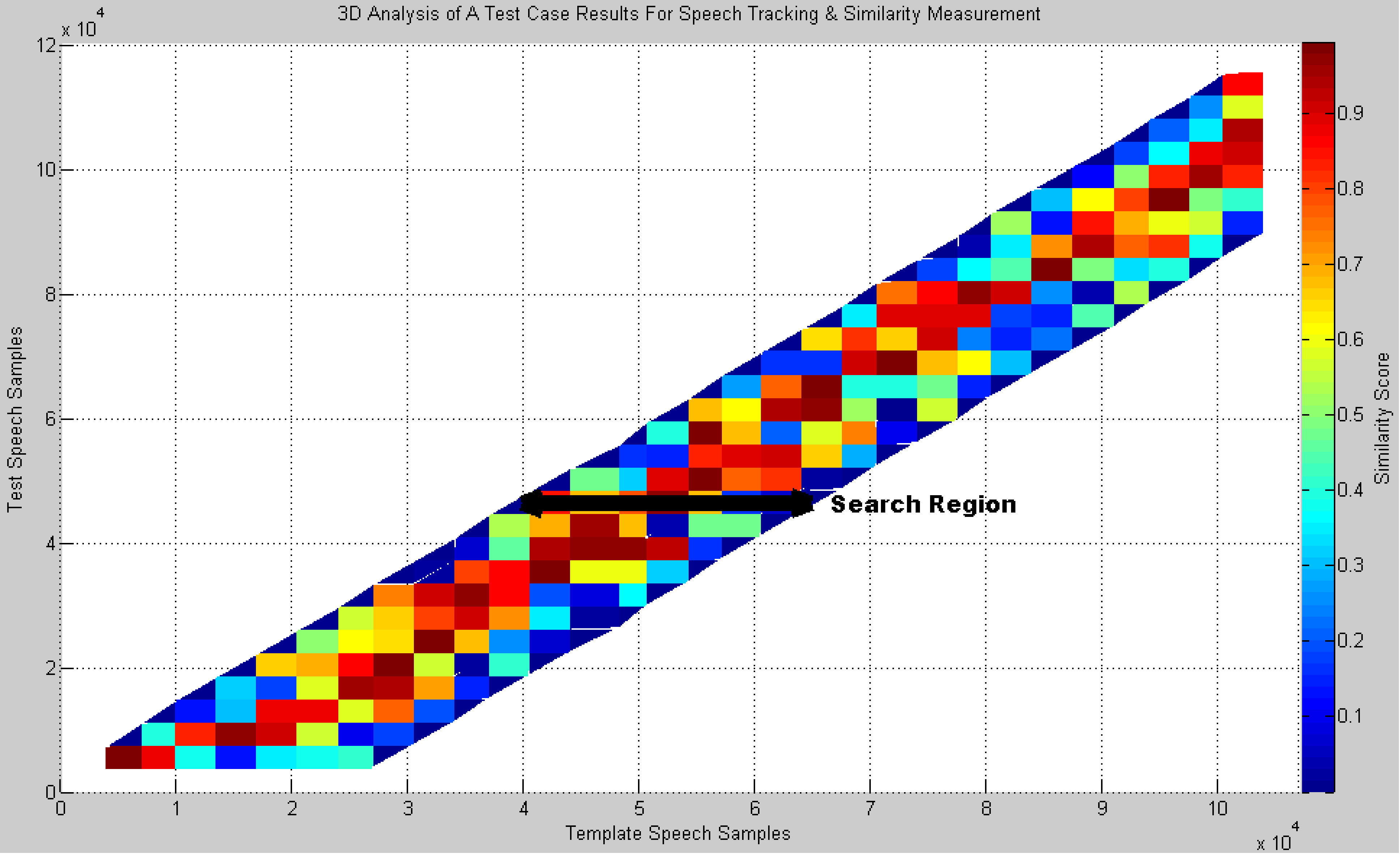


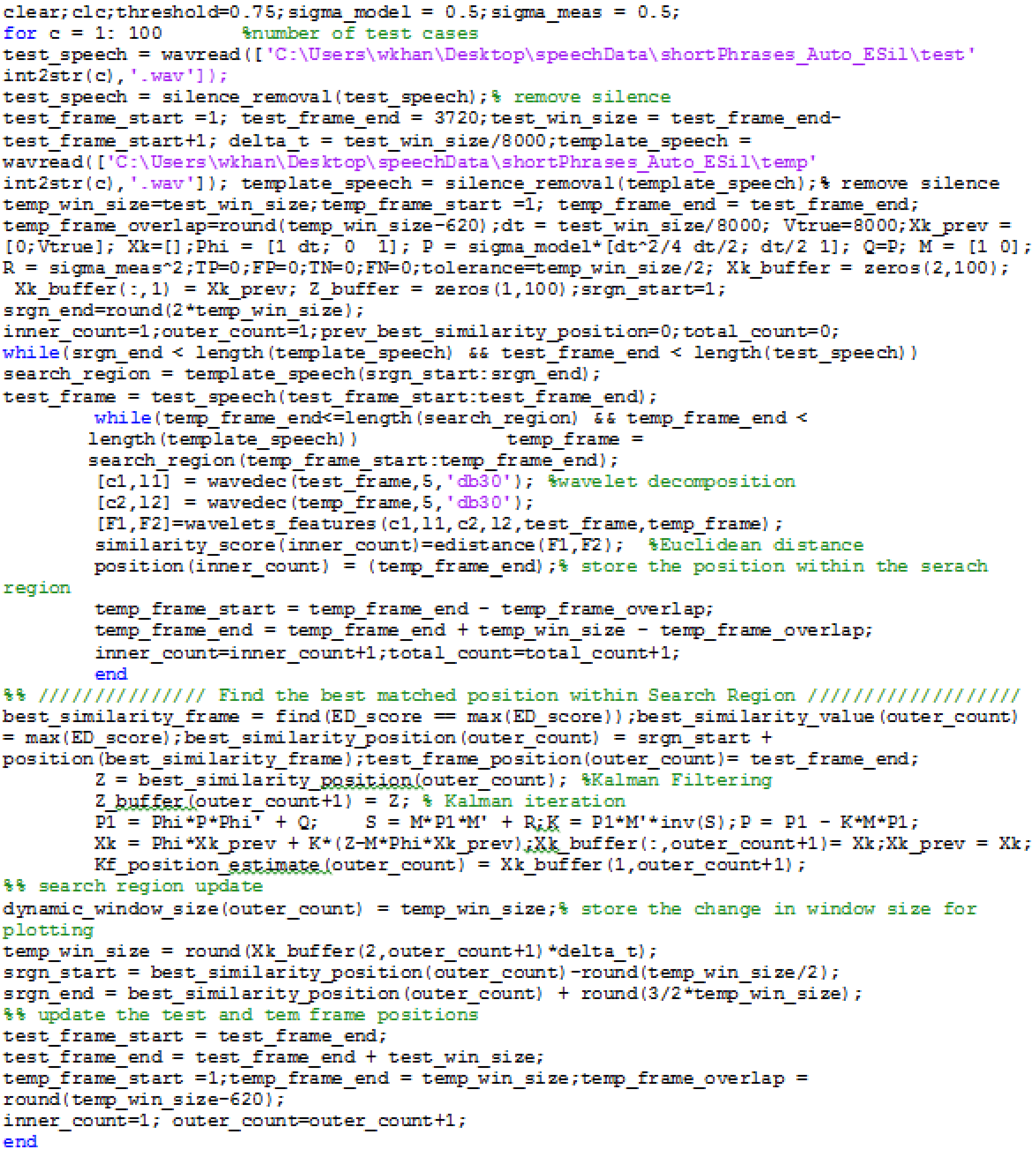
Figure 5-11: 3D Representation of Frame Size and Search Region Adaptation in TWCST for A Test Case

to consistency in the tracking path. It can also be analysed that the proposed approach resolves the time warping issue in efficient way. For 120,000 samples of test speech as compared to 10,6000 samples for template speech, TWCST approach provides a robust similarity matches between test and template frames without lsoing the tracking path.

Table 5-1: Accumulative Performance Comparison for Wavelet Based TWCST

Wavelet Based Adaptive Speech Tracking & Similarity Measurement Performance				
Evaluation Metrics		**Wavelet + Adaptive KF**	**Wavelet + Non-Adaptive SR**	**Wavelet + Adaptive KF (E & Spectral Centroid)**
Sensitivity		0.9907	0.9721	0.9951
Specificity		0.9841	0.9893	0.9654
Matching Accuracy		0.9869	0.9877	0.9764
1/LR+		0.0106	0.0108	0.0349
LR-		0.0107	0.0280	0.0052
F-Score		0.9833	0.9676	0.9683
Tracking Accuracy (%)		0.9983	0.9885	1
Type I Error	**μ**	0.0019	0.0010	0.0040
	σ	0.0070	0.0053	0.0104
Type II Error	**μ**	0.0086	0.0157	2.7232e-04
	σ	0.0460	0.0817	0.0011

Table 5-1 demonstrates the statistical results for wavelet based TWCST using multiple scenarios under diverse circumstances. Performance difference between KF based adaptive framing and search region based non-adaptive approach is presented using the gold standard validation metrics addressed in Table 4-2. The performance efficiency of the proposed approach is because of the adaptive framings that resolves the issue of dynamic nature of speech in terms of time warping. Also, the use of KF and DSM provides the substitute tracking information that never loses the tracking path when a mismatch or false positive occurs. This proves the reliability of the proposed approach for TWCST. Further detailed statistical results for all aforementioned scenarios using the dataset presented in Table 4-3 are addressed in Appendix B.



```
clear;clc;threshold=0.75;sigma_model = 0.5;sigma_meas = 0.5;
for c = 1: 100       %number of test cases
test_speech = wavread(['C:\Users\wkhan\Desktop\speechData\shortPhrases_Auto_ESil\test'
int2str(c),'.wav']);
test_speech = silence_removal(test_speech);% remove silence
test_frame_start =1; test_frame_end = 3720;test_win_size = test_frame_end-
test_frame_start+1; delta_t = test_win_size/8000;template_speech =
wavread(['C:\Users\wkhan\Desktop\speechData\shortPhrases_Auto_ESil\temp'
int2str(c),'.wav']); template_speech = silence_removal(template_speech);% remove silence
temp_win_size=test_win_size;temp_frame_start =1; temp_frame_end = test_frame_end;
temp_frame_overlap=round(temp_win_size-620);dt = test_win_size/8000; Vtrue=8000;Xk_prev =
[0;Vtrue]; Xk=[];Phi = [1 dt; 0  1]; P = sigma_model*[dt^2/4 dt/2; dt/2 1]; Q=P; M = [1 0];
R = sigma_meas^2;TP=0;FP=0;TN=0;FN=0;tolerance=temp_win_size/2; Xk_buffer = zeros(2,100);
 Xk_buffer(:,1) = Xk_prev; Z_buffer = zeros(1,100);srgn_start=1;
srgn_end=round(2*temp_win_size);
inner_count=1;outer_count=1;prev_best_similarity_position=0;total_count=0;
while(srgn_end < length(template_speech) && test_frame_end < length(test_speech))
search_region = template_speech(srgn_start:srgn_end);
test_frame = test_speech(test_frame_start:test_frame_end);
        while(temp_frame_end<=length(search_region) && temp_frame_end <
       length(template_speech))               temp_frame =
       search_region(temp_frame_start:temp_frame_end);
        [c1,l1] = wavedec(test_frame,5,'db30'); %wavelet decomposition
        [c2,l2] = wavedec(temp_frame,5,'db30');
        [F1,F2]=wavelets_features(c1,l1,c2,l2,test_frame,temp_frame);
        similarity_score(inner_count)=edistance(F1,F2);  %Euclidean distance
        position(inner_count) = (temp_frame_end);% store the position within the serach
region
        temp_frame_start = temp_frame_end - temp_frame_overlap;
        temp_frame_end = temp_frame_end + temp_win_size - temp_frame_overlap;
        inner_count=inner_count+1;total_count=total_count+1;
        end
%% /////////////// Find the best matched position within Search Region //////////////////
best_similarity_frame = find(ED_score == max(ED_score));best_similarity_value(outer_count)
= max(ED_score);best_similarity_position(outer_count) = srgn_start +
position(best_similarity_frame);test_frame_position(outer_count)= test_frame_end;
        Z = best_similarity_position(outer_count); %Kalman Filtering
        Z_buffer(outer_count+1) = Z; % Kalman iteration
        P1 = Phi*P*Phi' + Q;    S = M*P1*M' + R;K = P1*M'*inv(S);P = P1 - K*M*P1;
        Xk = Phi*Xk_prev + K*(Z-M*Phi*Xk_prev);Xk_buffer(:,outer_count+1)= Xk;Xk_prev = Xk;
        Kf_position_estimate(outer_count) = Xk_buffer(1,outer_count+1);
%% search region update
dynamic_window_size(outer_count) = temp_win_size;% store the change in window size for
plotting
temp_win_size = round(Xk_buffer(2,outer_count+1)*delta_t);
srgn_start = best_similarity_position(outer_count)-round(temp_win_size/2);
srgn_end = best_similarity_position(outer_count) + round(3/2*temp_win_size);
%% update the test and tem frame positions
test_frame_start = test_frame_end;
test_frame_end = test_frame_end + test_win_size;
temp_frame_start =1;temp_frame_end = temp_win_size;temp_frame_overlap =
round(temp_win_size-620);
inner_count=1; outer_count=outer_count+1;
end
```

Figure 5-12: Matlab script for TWCST using WD

5.4. Comparison of Proposed Time Warped Continuous Speech Tracking Approaches with Existing Methodologies

As compared to existing similarity measure and time warping techniques, there are a number of advantages of the proposed research contributions for TWCST approaches that are presented in the following paragraphs.

A. Adaptive Frame Size

Speech signals are naturally time warped which means that each word (utterance) may have different length if spoken at different time. In the literature, DTW with different variations has been used as the best approach to deal with time warped signals and minimize the warping path between two time series data inputs (Cassisi 2012), (Chan and Lee 2010), (Thambiratmann and Sridharan 2007), (Zhang and Glass 2011), (Carlin et al. 2011), (Zhang and Glass 2010), (Jansen et al. 2010), (Zhang et al. 2012). However, in case of speech dynamics, each corresponding frame in the template speech may be time warped, and hence it might be helpful to use dynamic frame size that changes according to the speaking speed. This issue is resolved in the current research work (Khan and Holton 2015) and (Khan et al. 2014) by introducing the concept of an adaptive frame size in TWCST approach. Equation 4-17 shows how the template frame size changes recursively according to the speaking (input speech signal) speed. Also, statistical results in Table 4-4, Table 5-1, Figure 4-11, Figure 5-6, and Appendix B demonstrate the performance comparison between adaptive and static framing based speech tracking techniques.

B. Search Region

An important contribution is made in the current research study by introducing the concept of a search region (Khan et al. 2014; Khan and Holton 2015). A search region is the template speech segment that is always larger than the test frame as shown in Figure 4-9. The test speech signal proceeds along the template speech signal within the search region and the best matched position is picked up as represented in Equation 4-4. Without search region, the DTW and speech tracking techniques would only be able to map each corresponding frame of test and template speech which might produce false results that will lose the tracking path due to the time warping phenomenon. Moreover, the computation cost for traditional DTW will increase exponentially with respect to speech signal length as discussed in Chapter 3, section 3.4. Performance evaluation for search region based DTW is presented in Appendix B. It can be analysed that the computation cost of a DTW is sufficiently decreased using the mean MFCC feature set. Moreover, the similarity matching and speech tracking performances are also increased as compared to traditional mapping of the test and template frames.

C. Computational Cost

As mentioned earlier, most of the related research work is based on STD, QbyE, and isolated word/utterance matching where the DTW is applied with a number of variations to minimize the warping path. However in case of continuous speech signals longer than a couple of words, the traditional DTW approach can’t be utilised without the collaboration of search region. This is because of the exponential growth in the size of distance matrix (Cheng-Tao et al. 2014) as shown in Figure 3-4. As an example, for a 10 seconds long test and template speech signal recorded at 8 KHz sampling frequency, DTW would need to map 80,000 times distance measure.

Therefore, the distance matrix size will be 80,000 x 80,000. This issue can be resolved using the boundary constraints to prune the warping path as shown in Figure 3-7. However, this sacrifices the desired performance in sufficient amount as shown in statistical results in Table 4-4 and Appendix B. This issue is resolved in the current research work by introducing the idea of segmentation and search region based speech tracking (Khan and Holton 2015) and mean-MFCC feature vectors. Multiple scenarios are generated where the segmentation (framing) and search region are applied to DTW and the proposed TWCST and the performance is compared in terms of computational cost as well as for speech tracking and similarity measure.

D. Multiple Source of Information

One of the major contributions of the research work presented in (Khan et al. 2014), (Khan and Holton 2015) is the idea of a DSM where test speech signal is analysed frame by frame. Each test frame is considered as a linearly moving object along the template speech signal within the search region at current time as shown in Figure 4-9. A mathematical implementation of DSM is presented in Chapter 4, Section 4.3. The existing research work for keyword spotting and continuous speech matching uses a single source of information to make decision for match/mismatch of two speech signals. For example DTW uses Euclidean distance and cosine angle for calculating the degree of dissimilarity and matching score respectively. However, considering multiple resources and combining their beliefs supports the decision making. In the proposed speech tracking approach, an observation from Euclidean distance based similarity match is produced which is fused with the DSM estimate using a KF. The KF processes both inputs and provides a combined position estimate at current time. Another advantage of including DSM is the recovery in case of losing the tracking path. The DTW approach will provide sudden surges in case of mismatches; consequently lose the tracking path. However, in the proposed approach, a continuous estimate is produced by a DSM which supports the feature based observation in case of a mismatch; therefore resulting a smooth tracking path. The tracking accuracy for all aforementioned approaches is presented in the statistical results (Table 4-4, Table 5-1).

E. Noise Covariance

There is always an uncertainty in the model (process) which indicates the error in process and the aim is to minimise this error. The beauty of KF is that it recursively updates the states according to noise covariance at each time step as shown in Equation 4-15. In the current approach for speech tracking (Khan and Holton 2015), noise variance plays multiple roles. Firstly, it measure the quantity of error that exists in terms of process and measurement noise as represented in Equation 4-15. Secondly, the noise variance provides the degree of dependency on both observations. That means how much the model depends upon the DSM and Euclidean distance position estimate. In other words, noise covariance assigns initial weights to each observation on the basis of which a KF tunes itself recursively as shown in Figure 4-8. On the other hand, to the best of our knowledge, existing approaches for speech similarity measure and DTW techniques don’t use noise variances.

5.5. Summary

In this chapter, a comprehensive overview of the research contribution towards wavelet based dynamic noise filtration is presented. A comparative study is conducted for wavelet based TWCST approach and previous approach as presented in Chapter 4. The formulation of WD and its application for noise filtration are introduced. For the first time, the concept of time-frequency domain noise filtration in the speech signal is applied for TWCST purpose. Likewise the previous approach, a DSM is used that is based on object's linear motion. A detailed flowchart for sequential processing of the wavelet based TWCST is presented. As discussed in Chapter 4, the deployment of KF and DSM is novel in speech tracking in many aspects. It empowers not only the decision of match/mismatch but also provides a backup to recover the tracking path in case of mismatch or false negatives. Multiple speech corpuses are used for evaluating the performance for newly introduced TWCST approach using the gold standards metrics for binary classifier (i.e. match/mismatch). In addition to speech tracking, the new approach is compared with the existing work in terms of similarity matching and computational cost. Detailed statistical results are presented in form of tabular and graphical form. Finally, advantages of the proposed approaches for TWCST over the existing methodologies in terms of theoretical and practical aspects are discussed.

6. KEY-WORD SPOTTING IN CONTINUOUS SPEECH

6.1. Scope

This chapter consists of five sections that present a comprehensive overview of keyword spotting also known as Spoken Term Detection (STD) in continuous speech. First section introduces the idea of keyword spotting and its practical applications. Formulation of the Dempster-Shafer's theory of mass combination and its application to keyword spotting (Khan and Holton 2015) is addressed in Section 6.3. For the first time, the theory of belief combination is used in the speech related work. The idea is to combine the similarity beliefs produced by more than one resource to make a final decision (belief) of keyword match/mismatch. These resources are in the form of distance metrics and feature set. A detailed discussion on the implementation of newly introduced keyword spotting followed by a comprehensive discussion on the comparison with existing approaches is also presented in this section. Section 6.4 presents the research contribution (Khan et al. 2012) in the form of a PPM deployment for word identification in the presence of background noise. The mathematical formulation and performance analysis is discussed in detail. Section 6.5 presents the contributed work (Khan et al. 2013) towards the introduction of a similarity measure approach that is based on vector addition method. Its mathematical justification and relationship with other distance metrics is presented. Finally, a summary of the chapter is presented in the last section.

6.2. Introduction

Keyword spotting in continuous speech is an emerging but challenging task that needs to deal with speech dynamics. Literature consists a number of keyword spotting approaches in relation to Query By Example (QbyE) (Anguera et al. 2014), (Tejedor et al. 2013), and STD (Anguera et al. 2013), (Chan and Lee 2013) that use some sort of variations in DTW (Zhang and Glass 2009; 2011), (Chunan and Lin-shan 2010). Over the past decade, most of the related research is focused on novelty of template representation methods (Fousek and Hermansky 2006), (Hazen et al. 2009), (Huijbregts et al. 2011). A detailed review of the existing methodologies for keyword spotting is presented in Chapter 3, section 3.5.2. In the proposed research, a valuable contribution to the existing keyword spotting approaches is made in the form of dynamic noise filtration method that uses the WD to filter out the unnecessary time domain as well as frequency domain components from the speech signal (Khan and Holton 2014). Frequency components with maximum correlation with mother wavelet are forwarded for the feature extraction and signal matching process.

In addition, for the first time, the Dempster-Shafer's Theory (DST) of combined evidence is deployed for keyword spotting purpose (Khan and Holton 2015). A most interesting aspect of DST is the combination of beliefs (i.e. probabilities) obtained from multiple resources and the modelling of conflict between them. In case of keyword spotting, DST can play a significant role while integrating the keywords match/mismatch beliefs (i.e. scores) from multiple resources and providing a final combined belief. The evidence resources may include a variety of distance metrics, extracted features, and/or keyword spotting approaches. Practically, there are numerous occasions where one resource provides better belief than the other. Therefore, it is helpful to combine the evidences from multiple approaches and make the final decision about the keyword (i.e. target word) occurrence in the continuous speech. Keyword spotting approach plays an important role in practical life. As an example, a keyword spotter can be useful for localization of the specific word occurrences in a long speech (e.g. telephone) recording that would be useful for the intelligence agencies and security organisations. Similarly, wake-up word to activate a device or initiate the voice recognition platform is another area of application.

6.3. Key-word Spotting and Dempster-Shafer's Theory of Evidence

Keyword spotting can be considered as a sub-part of the ASR which aims to extract the partial information from speech signal in the form of query utterance (keyword). Figure 6-1 shows the keyword spotting task addressed in the proposed research study where the keyword would be identified regardless of the spoken language.

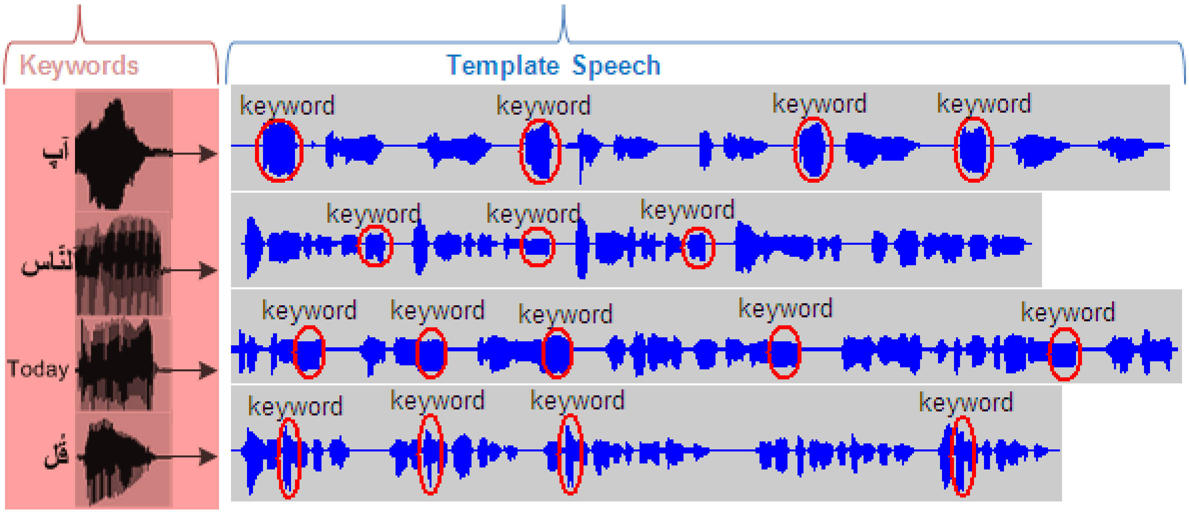
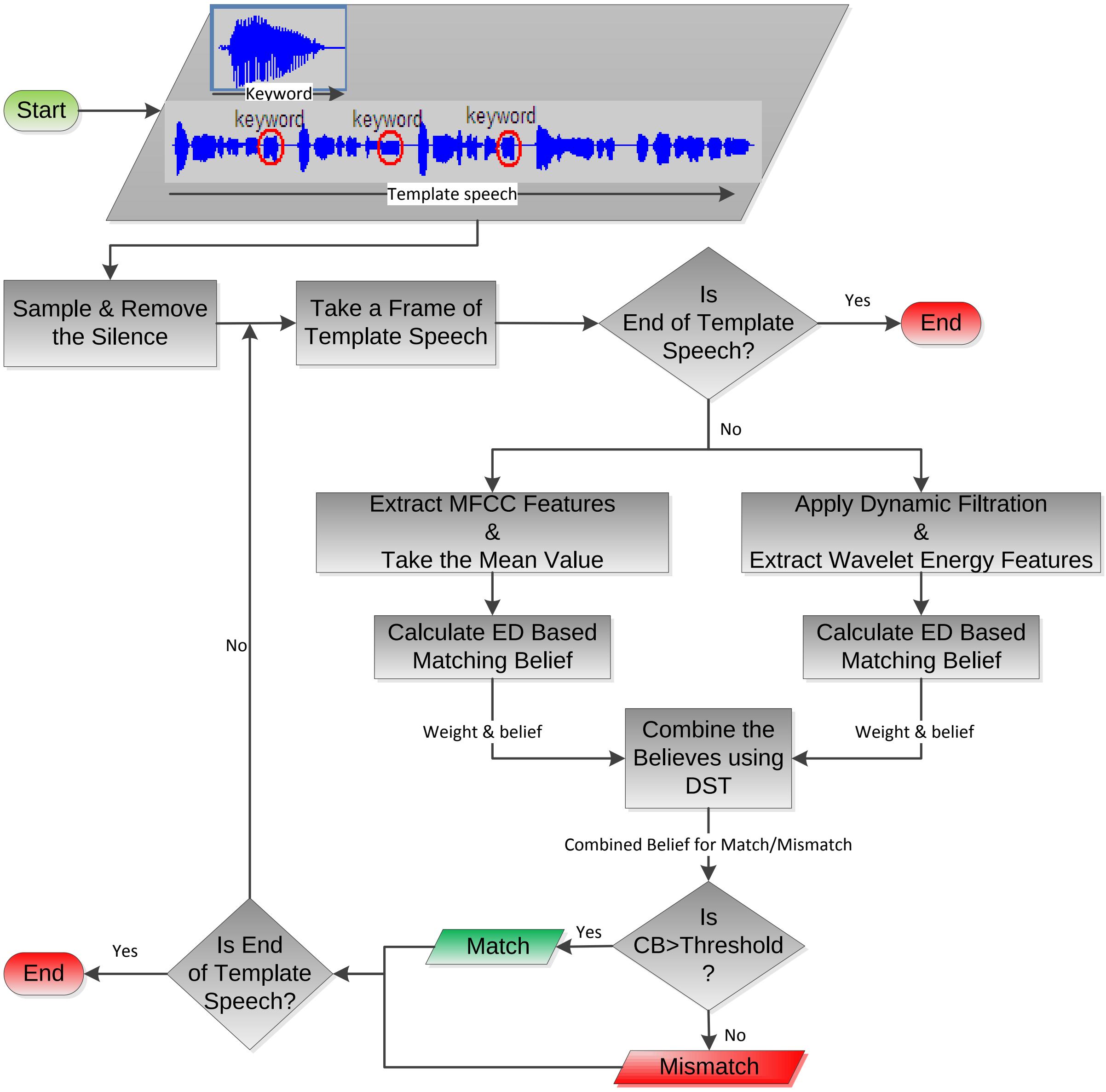


Figure 6-1: Keyword Spotting in Continuous Speech

As discussed in Section 3.5.2, one of the major challenges associated with the existing approaches is the time warping. Due to dynamic length of spoken words and existence of the silence segments, performance of the existence techniques degrades. To resolve the speech dynamics and time warping issues for the proposed keyword spotting, a number of methodologies are amalgamated sequentially as shown in Figure 6-2.



Start
Keyword
keyword
keyword
keyword
Template speech
Sample & Remove the Silence
Take a Frame of Template Speech
Is End of Template Speech?
Yes
End
No
Extract MFCC Features & Take the Mean Value
Apply Dynamic Filtration & Extract Wavelet Energy Features
Calculate ED Based Matching Belief
Calculate ED Based Matching Belief
Weight & belief
Weight & belief
Combine the Believes using DST
Combined Belief for Match/Mismatch
Is CB>Threshold ?
Yes
Match
No
Mismatch
Is End of Template Speech?
Yes
End
No

Figure 6-2: Processing Flow of the Keyword Spotting

Firstly, the template and keyword signals are processed by a silence removal technique to filter out unnecessary segments. Enhanced signals are then forwarded to the framing process. In the next step, the most dominant acoustic features that represent the speech frames are extracted and forwarded for the similarity measurement. Finally, the similarity scores (i.e. beliefs) from multiple similarity measures are combined using DST that provides a combined belief for keyword match/mismatch. Further discussion and mathematical formulation of each component in Figure 6-2 is presented in the following sections.

6.3.1. Pre-Processing

In the first step, the template speech and keyword utterances are forwarded to the pre-processing unit which enhances the speech quality in terms of resampling and silence removal. A detailed formulation for sampling and silence removal is already discussed in Chapter 2, section 2.2.1. There are a number of methodologies in the literature that have been used for silence removal as presented in Chapter 2. However, an efficient approach is introduced in the proposed research (Khan and Holton 2015) that is based on pitch tracking. Detailed implementation of the proposed silence removal approach is addressed in Chapter 4, Section 4.3.

6.3.2. Speech Signal Framing

In second step, speech framing is applied to the enhanced template speech that recursively crop a fixed length frame and forward it for further processing until the end of template speech. In order to handle the occurrence of discontinuities at the frame boundaries, keyword is matched to the overlapping frames of template speech. Figure 6-3 demonstrates the keyword progression along the overlapped frames of template speech. A detailed mathematical formulation of framing process and related work is presented in Chapter 2, section 2.2.1.

6.3.3. Dynamic Speech Filter and Feature Extraction

The segmented speech frames for keyword and template speech are forwarded for the feature extraction process. Rather than extracting only MFCC features from the speech signal, wavelet based energy features are also extracted. Combination of the MFCC mean values and wavelet energy based features empowers the keyword spotting performance that is discussed later in the performance section. A detailed implementation of the MFFC feature extraction is presented in Section 2.2.2.

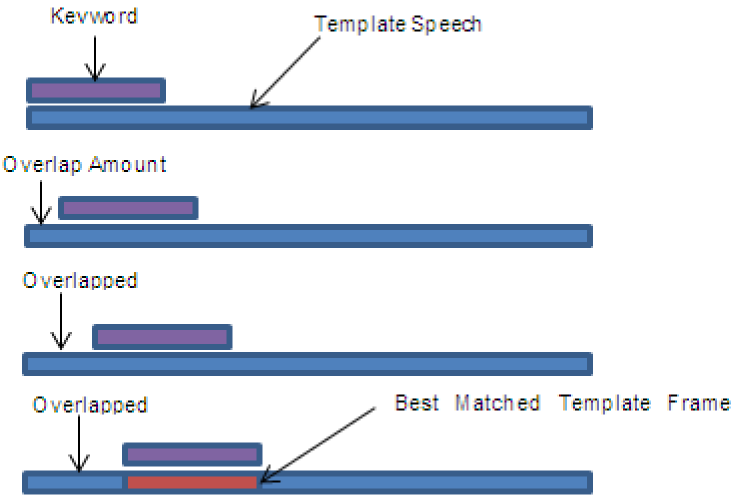


Figure 6-3: Progression of Keyword along Template Speech

Also the process of dynamic noise filtration and wavelet based feature extraction is presented in Chapter 5, section 5.2. A major advantage of WD is the simultaneous representation of time-frequency components which helps to filter out the noisy segments in the spectrum as well as from the time-domain speech signal. Figure 6-4 presents the block diagrams for MFCCs and WD based energy features.

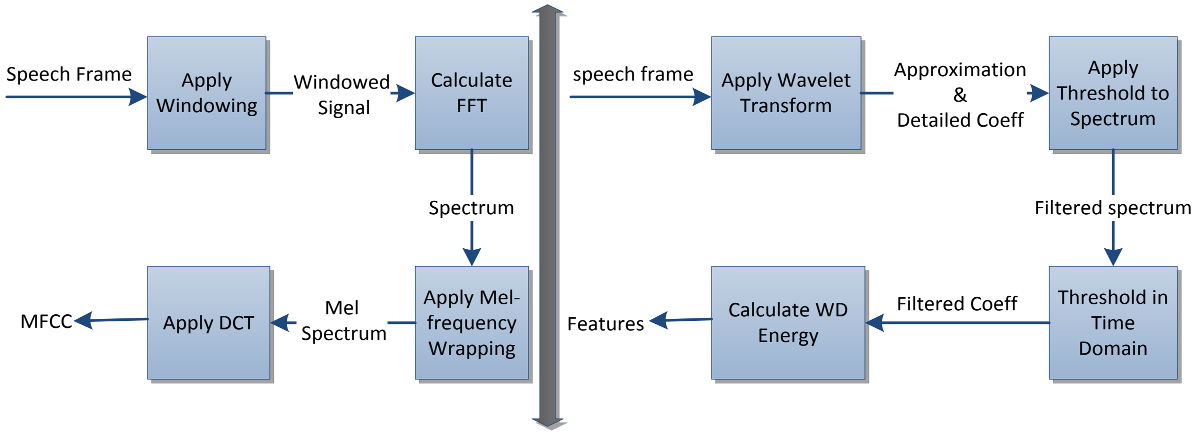


Figure 6-4: Block Diagram for MFCC and Wavelet Features Extraction

6.3.4. Similarity Measurement

In the next step, the extracted MFFC features for both; keyword and template frames are normalised and forwarded to a similarity measure. A Euclidean distance is used as a similarity measure. As Euclidean distance provides dissimilarity score, fewer score means more similar. A detailed formulation and related work for Euclidean distance is presented in Chapter 3, section 3.3. Simultaneously, the wavelet energy based features are also forwarded to Euclidean distance and a similarity score along with template frame position is measured. The extracted feature vectors are normalised to always get a similarity scores between zero and one.

Suppose the mean MFCC feature vectors for keyword and template speech frame are represented as $km = \{km_{1,} km_{2,} ....km_n\}$ and $tm = \{tm_{1,} tm_{2,} ....tm_n\}$ respectively. Similarly, $ke = \{ke_{1,} ke_{2,} ....\mathrm{k}e_n\}$ and $te = \{te_{1,} te_{2,} ....t\mathrm{e}_n\}$ represents the wavelet energy feature vectors for keyword and template speech frame respectively. As the keyword and template speech frames size is static, the extracted features consists the same length and is denoted by 'n'. The normalised Euclidean based similarity measure for aforementioned features can be defined as:

$$S_m = 1 - \sum_{i=1}^{n} \sqrt{\left(\frac{km_i}{|km|} - \frac{tm_i}{|tm|} \right)^2} \qquad 6\text{-}1$$

$$S_e = 1 - \sum_{i=1}^{n} \sqrt{\left(\frac{ke_i}{|ke|} - \frac{te_i}{|te|} \right)^2} \qquad 6\text{-}2$$

Where S_m and S_e are the similarity measurements between keyword and template speech frames for MFCC and wavelet energy based features respectively. Because of the normalised data distribution, the similarity measurement values are within the ranges of 0 and 1 for each pair of keyword and template speech frame. These measurements are used as the match/mismatch beliefs that are forwarded to DST approach for beliefs combination along with the pre-set weights.

6.3.5. Formulation of the Combination of Beliefs

The DST is considered as a generalization to the Bayesian theory in such a way that it can handle the degree of ignorance (Foley 2012). In case of certain information, there are a number of fusion methods that can provide the combined belief. However, most of these approaches are unable to handle the degree of ignorance. The DST provides the best estimate for the degree of belief by combination of evidences/ believes (CE) from multiple resources. Along with the advantages of the DST, it has been criticized by the researchers due to lack of ability to deal with high degree of conflict. A detailed study on DST advantages, disadvantages, criticism, and its application areas is presented in (Foley 2012). In the proposed study, for the first time the DST is used for evidence combination in the keyword spotting. The advantage of this method is to obtain the basic probability assignment based on the similarity values obtained from Euclidean distance. Also, it empowers the decision making for keyword match/mismatch to be dependent on multiple resources simultaneously. Figure 6-5 shows the block diagram of DST for the proposed keyword spotting.

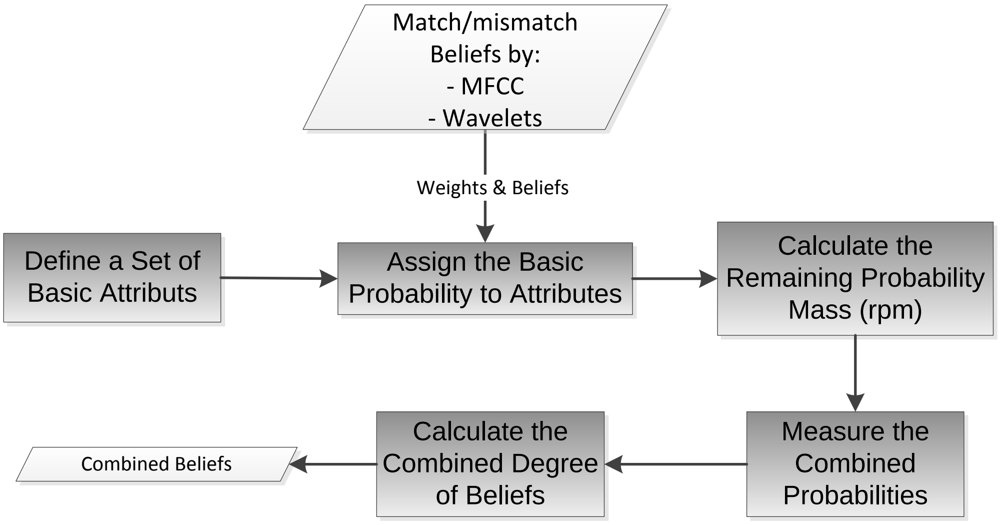

Match/mismatch
Beliefs by:
- MFCC
- Wavelets
Weights & Beliefs
Define a Set of Basic Attributs
Assign the Basic Probability to Attributes
Calculate the Remaining Probability Mass (rpm)
Measure the Combined Probabilities
Calculate the Combined Degree of Beliefs
Combined Beliefs

Figure 6-5: Block Diagram for Deployment of DST in Keyword Spotting

Mathematical formulation of the DST in terms of keyword spotting approach is represented in the following steps.

Define the Basic Attributes:

Let's $E = \{mel, e\}$represents the set of basic attributes for the proposed keyword spotting where 'mel' and 'e' are the belief resources. In our case, MFCC and wavelet energy based features are such resources. The relative weights for the basic attributes are pre-set by offline experiments (as discussed in performance analysis section) such that $0 \le \omega_i \le 1$ and it fulfils Equation 6-3.

$$\sum_{i=1}^{L} \omega_i = 1 \tag{6-3}$$

Where; '*L=2*' represent the number of attributes that are (mel, e) for keyword spotting.

The distinctive evaluation grades are defined as set of two entities, i.e.

$$H = \{match, mis\_match\} \tag{6-4}$$

For each attribute in 'E' and evaluation grade 'H', a degree of belief β_n is assigned. The degree of belief denotes the source's level of confidence when assessing the level of fulfilment of a certain property.

Basic Probability Assignments for Each Basic Attribute:

Let $m_{n,i}$ be a basic probability mass representing the degree to which the i^{th} basic attribute. A hypothesis that the general attribute is assessed to the n^{th} evaluation grade H_n can be presented as:

$$m_{n,i} = \omega_i \beta_{n,i} \qquad 6\text{-}5$$

Where 'n' are the number of evaluation grades (i.e. $match, mis\_match$). The remaining probability mass $m_{H,i}$ unassigned to each basic attribute is calculated as:

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{2} \beta_{n,i} \qquad 6\text{-}6$$

'N =2' are the total number of evaluation grades.

The remaining probability mass is further decomposed into $\bar{m}_{H,i}$ and $\tilde{m}_{H,i}$ as:

$$\bar{m}_{H,i} = 1 - \omega_i \qquad 6\text{-}7$$

$$\tilde{m}_{H,i} = \omega_i \left(1 - \sum_{n=1}^{2} \beta_{n,i} \right) \qquad 6\text{-}8$$

With

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \qquad 6\text{-}9$$

Equation 6-7 measures the degree to which final attributes have not been assessed yet to individual grades due to the relative importance of basic attributes after their aggregation. Equation 6-8 measures the degree to which final attributes cannot be assessed to individual grades due to the incomplete assessments for basic attributes.

Combined Probability Assignments:

In this step, the probability mass of the basic attributes $E = \{mel, e\}$ are aggregated to form a single assessment for keyword match/mismatch. The combined probability masses can be generated using the following set of recursive evidence reasoning equations:

$$\{H_n\}:$$
$$m_{n,i+1} = K_{i+1}[m_{n,i}.m_{n,i+1} + m_{H,i}.m_{n,i+1} + m_{n,i}.m_{H,i+1}] \qquad \text{6-10}$$
$$n = 1, \ldots, N$$

Where $i = \{1, \ldots, L-1\}$, L =2 are the number of basic attributes, and 'N =2' are the total number of evaluation grades.

In above equation, $m_{n,1}.m_{n,2}$ measures the degree of both attributes $\{mel, e\}$ supporting the general attribute of keyword match to be assessed to H_n. The term $m_{n,1}.m_{H,2}$ measures the degree of only 1st attribute $\{mel\}$ supporting keyword match to be assessed to H_n. The term $m_{H,1}.m_{n,2}$ measures the degree of only 2nd attribute $\{e\}$ supporting final belief to be assessed to H_n.

$\{H\}$:

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \qquad \text{6-11}$$

$$\tilde{m}_{H,i+1} = K_{i+1}[\tilde{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i+1}.\tilde{m}_{H,i}] \qquad \text{6-12}$$

$$\bar{m}_{H,i+1} = K_{i+1}[\bar{m}_{H,i}.\bar{m}_{H,i+1}] \qquad 6\text{-}13$$

$$K_{i+1} = \left[1 - \sum_{t=1}^{N=2} \sum_{\substack{j=1 \\ j \neq t}}^{N=2} m_{t,i}.\ m_{j,i+1} \right]^{-1} \text{, for } i = \{1,...,L-1\} \qquad 6\text{-}14$$

In Equation 6-12, $\tilde{m}_{H,1}.\tilde{m}_{H,2}$ measures the degree to which final attribute cannot be assessed to any individual grades $\{match, mis\_match\}$ due to the incomplete assessments for both attributes$\{mel, e\}$. Term $\bar{m}_{H,1}.\tilde{m}_{H,2}$ measures the degree to which final attributes cannot be assessed due to the incomplete assessments for $\{mel\}$ only. In Equation 6-13, $\bar{m}_{H,1}.\bar{m}_{H,2}$measures the degree to which final attributes have not been assessed yet to individual grades due to the relative importance of $\{mel\}$ and $\{e\}$after $\{mel\}$ and $\{e\}$have been aggregated. The normalization factor $'K'$ is used to normalize m_n, m_Hsuch that$\sum_{n=1}^{N=2} m_n + m_H = 1$.

Calculation of the Combined Degrees of Belief:

Let β_n denotes the combined degree of belief that a keyword spotting is assessed to the gradeH_n, which is generated by combining the assessments for all the associated basic attributes$E = \{mel, e\}$, thenβ_n is calculated by:

$$\{H_n\} : \beta_n = \frac{m_{n,L}}{1 - \bar{m}_{H,L}} \qquad n = 1,..,N \qquad 6\text{-}15$$

$$\{H\} : \beta_H = \frac{\tilde{m}_{H,L}}{1 - \bar{m}_{H,L}} \qquad 6\text{-}16$$

Above equation for $\beta_H$measures the belief that is left unassigned during the assessments.

6.3.6. Performance Evaluation and Experimental Setup

A number of metrics have been used in the literature for evaluating the performance of keyword spotting approaches. However, most relevant are the gold standards used for the binary classification (Soluade 2010). This is because of the output from keyword spotting approach is in the binary form (i.e. match or mismatch). Table 4-2 presents the detailed metrics that are used for the validation of the proposed keyword spotting approach.

True Positive (TP): The recognizer correctly spots a query (keyword) utterance with high confidence.

True Negative (TN): The recognizer correctly rejects an out-of-grammar keyword.

False Positive (FP): The recognizer incorrectly spots the query utterance.

False Negative (FN): The recognizer incorrectly rejects an in-grammar query utterance.

If there are ‘N’ utterances in template speech, then TP + FP + TN + FN = N.

Sensitivity (True Positive Rate): Probability that a keyword is positively matched with high confidence when the keyword is actually in-grammar (i.e. template frame). This is expressed as a percentage of all the in-grammar matches.

Specificity (True Negative Rate): Probability that a keyword is matched as out-of-grammar when it is indeed out-of-grammar and is therefore not accepted by the recogniser. This is expressed as a percentage of the of all the out-of-grammar matches.

Accuracy: This is a percentage of all the matches that were correctly classified.

Positive Likelihood Ratio: The ratio of the probability of positively recognising a keyword with high confidence when an in-grammar keyword is spoken, and the probability of positively recognising a keyword with high confidence when an out-of-grammar keyword is spoken. This is basically the True Positive Rate/False Positive Rate.

Negative Likelihood Ratio: The ratio of the probability of rejecting an in-grammar keyword and the probability of rejecting an out-of-grammar keyword.

F-Score: As described in Chapter 4, section 4.4.

Receiver Operating Characteristics (ROC): As described in Chapter 4, section 4.4.

The proposed keyword spotting approach is tested on two types of dataset. A case study is conducted on a recorded speech dataset by a number of speakers. This data is recorded in noise free lab environment using an efficient microphone that consist a built-in noise filter. Speakers from different gender, age, and accent recorded the template speeches and corresponding keywords. Table 6-1 demonstrates the keyword list, number of occurrences, number of speech recordings, and gender information. In addition, a number of dataset are considered that are available online. Details of these dataset are already presented in Chapter 4, section 4.4. For recording purpose, the SENNHEISER e935 is used which is a vocal dynamic microphone that consists a built in noise filter. Speech is recorded at a sampling frequency of 8KHZ. A laptop device with Intel® Core™ i5 CPU, 4 G-byte memory, 32 bit operating system and running the Window 7 home premium operating system is used for the processing and experimentation purpose.

6.3.7. Experimental Results

Detailed experiments are conducted using aforementioned metrics to evaluate the performance in terms of sensitivity, specificity, accuracy, likelihood ratios, absolute error, execution time, and F-score. Because of the template frames overlapping, a mismatch tolerance of one frame size is set throughout the experiment conduction.

In addition to traditional test validation methods, a number of important metrics are added that have mostly been used in the area of binary classification.

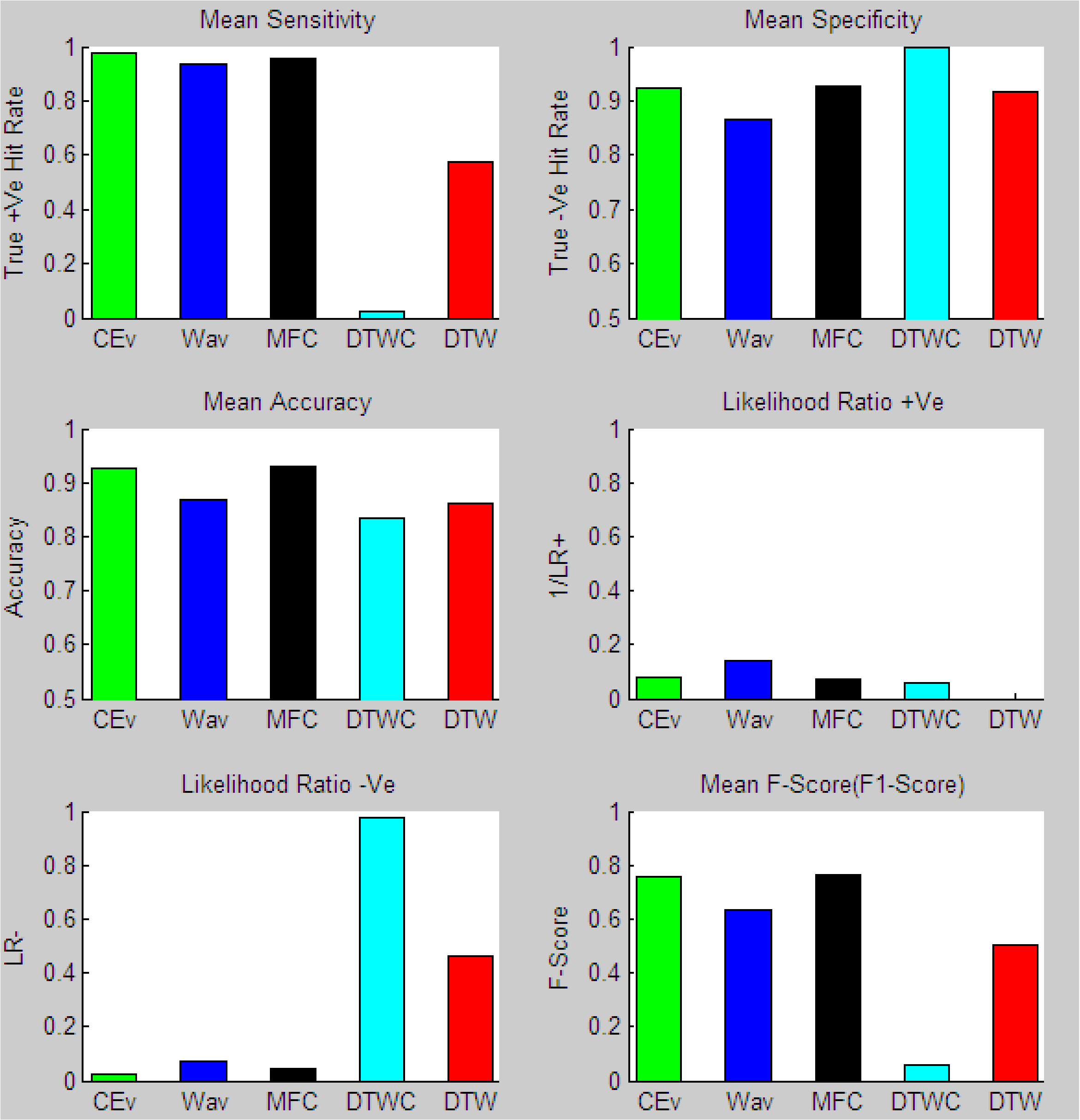


Figure 6-6: Results Comparison and Validation Using Different Metrics

Figure 6-6 demonstrates a detailed performance comparison of the proposed keyword spotting approach and other state-of-the-art keyword spotting approaches. Individual performances of the proposed keyword spotting approaches using combined evidence (CEv), wavelets (Wav), and MFCCs (MFC) are compared with the existing constrained DTW and enhanced DTW (i.e. segmented DTW) based keyword spotting approaches. It can be observed that the sensitivity of CEv is greater than the individual values of MFCCs and wavelets by a factor of 2% and 5% respectively. This implies that the deployment of DST increases the keyword spotting as well as it empowers the performance in terms of decision making. It can also be observed from the above figure that there is a dramatic decrease (50%) in the sensitivity when constrained is applied to conventional DTW. Despite of the fact that the search space in DTWC is far less than segmented DTW (as discussed in Chapter 3, section 3.4), yet the segmented DTW is better than DTWC in terms of keyword spotting outcomes.

The likelihood ratios (LR+, LR-) are considered as one of the best metrics to measure the diagnostic accuracy. In terms of keyword spotting, LR presents the probability of a test with keyword match divided by the probability of the same test with keyword mismatch. A larger LR+ consist more information than the smaller LR+ whereas smaller LR- consists more information than a larger LR-. To simplify the LR values, a relative magnitude is considered by taking the reciprocal of LR+. It is analysed from the Figure 6-6 that the LR- for CEv approaches to zero (0.03) as compared to 0.2 for DTW and 0.9 for DTWC which indicates the robustness of the proposed keyword spotting approach. Similarly, F-score is a measure that considers both *precision* and *recall* to measure the system performance. Figure 6-6 demonstrates that the F-score of CEv and MFC based keyword spotting is 80% as compared to 57% of DTW based keyword spotting approach.

Threshold Setting

An important factor for the proposed keyword spotting is the threshold value that specifies the decision boundary for the keyword to be considered as a match or mismatch. The trade-off between sensitivity and specificity depends upon the threshold value change. The threshold value is directly proportional to specificity and may vary with respect to the application area. For example, keyword spotting use for the intelligent agencies must assign more importance to sensitivity to maximize the true positives by reducing the threshold value. This is because the priority in such cases will not to miss a keyword (e.g. blast, terror) that exists in a speech recording. To set up the threshold value for the proposed approach, the ROC curves are achieved for various methods while conducting experiments on the dataset acquired in the aforementioned case study.

To set a threshold value for match/mismatch decision boundary, an ROC is achieved by varying threshold from 0 to 1 with a lag of 0.01 as shown in Figure 6-7. It is observed that the best value in the ROC is achieved with a threshold value of 0.85 (85%). It means that the template frame will be rejected if its matching belief with the keyword is less than 85%. As there is a trade-off between sensitivity and specificity, threshold value is chosen while considering both metrics.

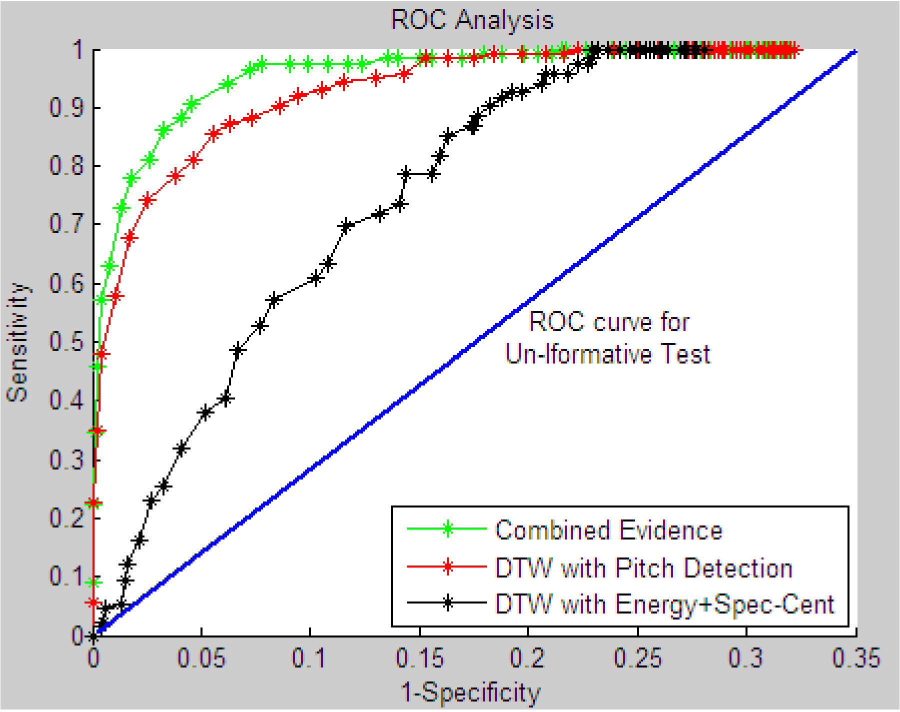


Figure 6-7: The ROC Curves for Varying Threshold Values for Match/Mismatch Decision

In addition to the best threshold value selection, the ROC curves in Figure 6-7 manifest the superiority of the CEv based keyword spotting as compared to state-of-the-art DTW approach. Another aspect of the ROC curves shown in Figure 6-7 is the validation of the silence removal approach introduced in this research study. It is clear that area under the curve for energy and spectral centroid based silence removal is far less than pitch detection based silence removal approach. Figure 6-8 shows a simultaneous representation of sensitivity, specificity, and threshold values that helps to analyse the relative variations in these metrics.

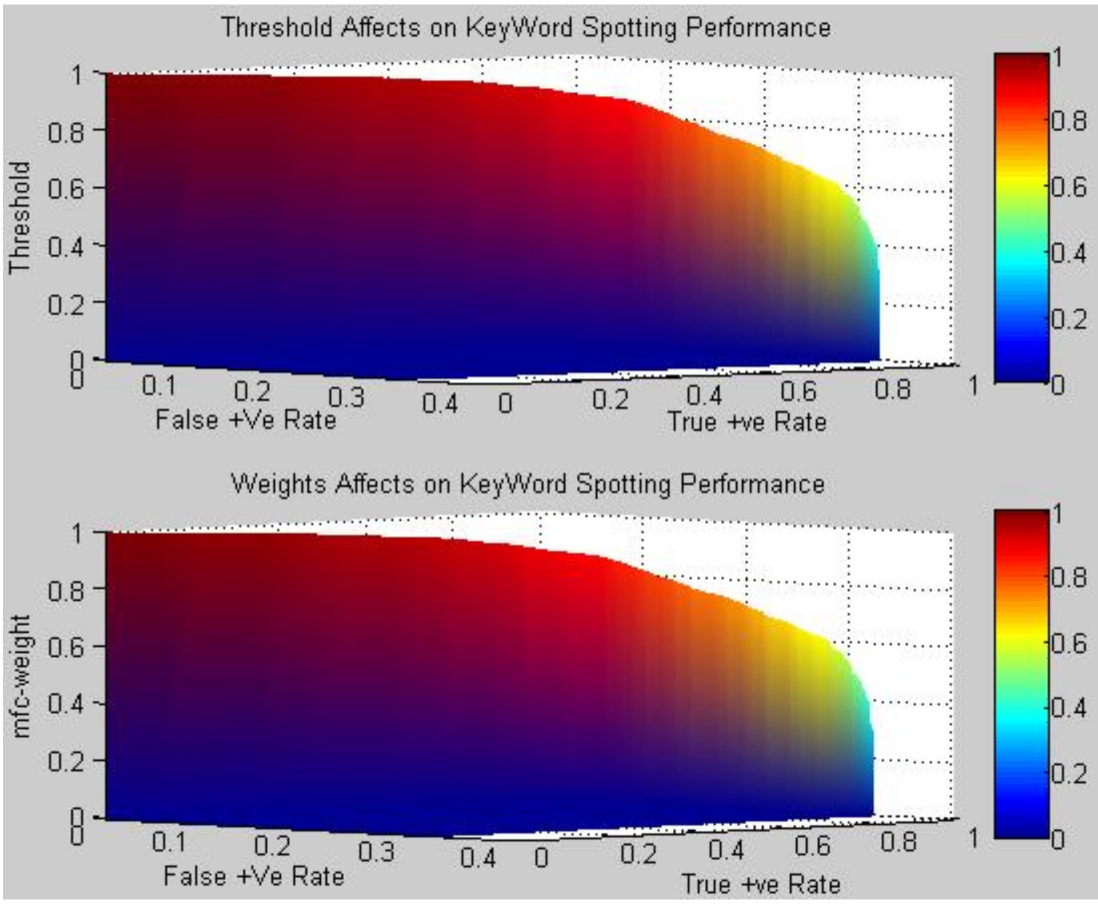


Figure 6-8: Setting the Weights to Belief Resources

Weight Allocation to Basic Attributes

It can be observed in Equation 6-7 and 6-8 that the final belief of keyword spotting depends upon the weights assigned to the basic attributes $E = \{mel, e\}$. Experiments are conducted by setting continuously varying weights for both attributes from 0 to 1 with a lag of 0.01. The ROC curve is achieved (Figure 6-8) for 100 values of weights between 0 and 1. It is observed that the best performance in terms of FPR and TPR is achieved at $w_{mfcc} = 0.75$ $w_{wav} = 0.25$. This implies that the best performance is achieved by assigning more weight to matching belief of MFCC. It is also clear from Figure 6-6, and Table 6-1 that the individual performance by both attributes $\{mel, e\}$ in terms of sensitivity and error rate is poor than the combined evidence. This validates the importance of DST for the keyword spotting.

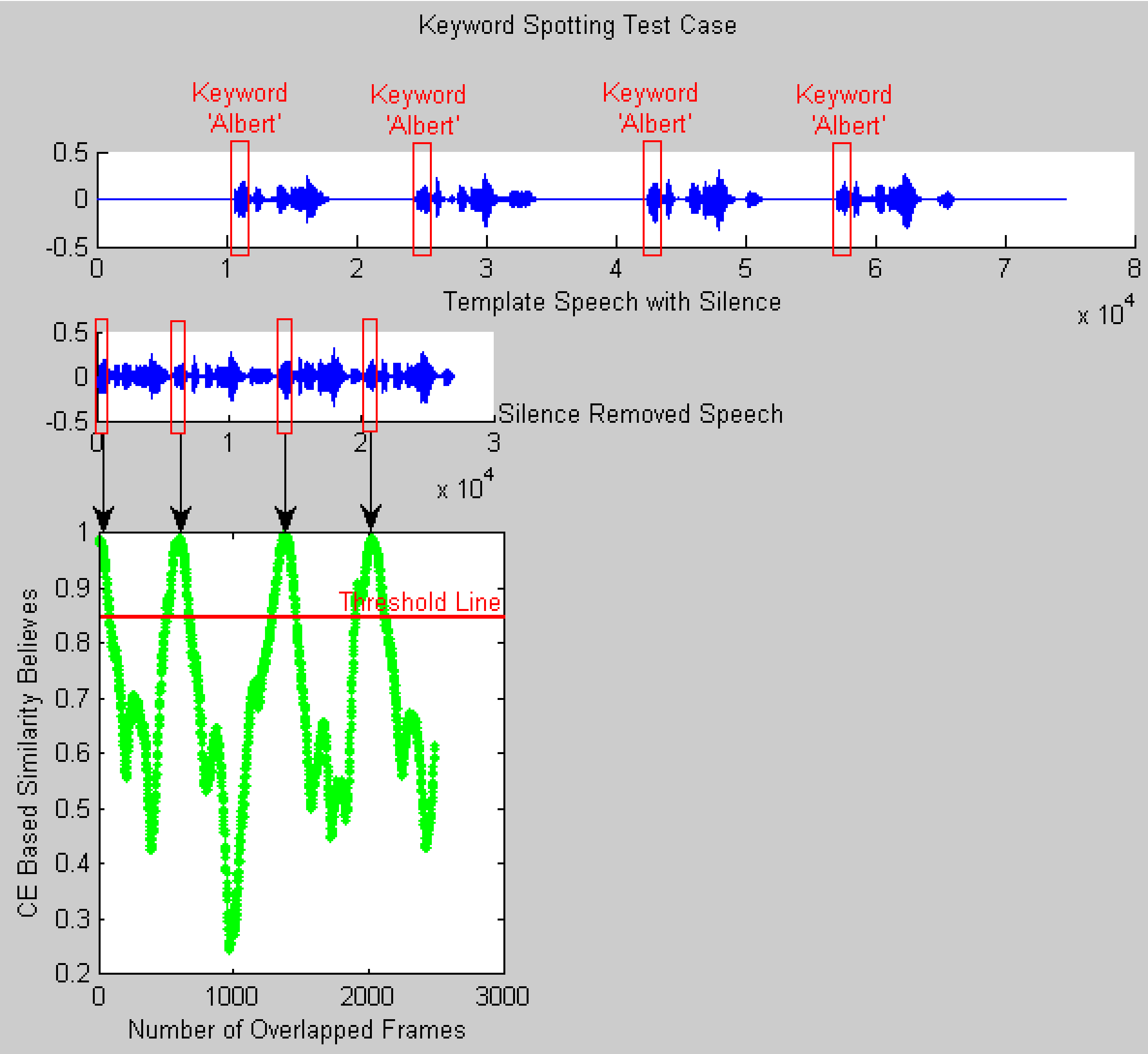


Figure 6-9: A Typical Keyword Spotting Example

Figure 6-9 demonstrates the proof of concept for a test case using a keyword {'Albert'}. The robustness of the proposed approach can be observed by mapping the spotted locations (i.e. peaks) in the template speech corresponding to the ground truth keyword positions. In addition, it can be easily observed in above figure that the length of template speech signal is significantly reduced by a significant amount (i.e. 60% approximately). This shows the robustness of the silence removal approach that is introduced very first time in this research (Khan and Holton 2015). Detailed experimental results for 35 mutilanguage keywords are presented in Table 6-1. Information about the datasetset is shown in terms of total number of occurences, length, gender, language, and keyword. As the proposed research study is based on the phonemes based signal processing, there is no training process involved leading to the language independency. It can be observed from the test metrics (TP, TN, FP, FN, sensitivity, specificity, accuracy, LR+, LR-) that the spotting performance is achieved consistently regardless of the spoken keywords.

Table 6-1: Performance Analysis of KeyWord Spotting Approaches using Confusion Matrix and Likelihood Ratios

Key-Word	No. of Rec	Key-Word/ Test Case	Speaker Gender	Combined Evidence (Mean-MFCC + Wavelets)									DTW + MFCC									DTW-Restricted + MFCC								
				TP	TN	FP	FN	Sen	Spec	Acc	1/LR+	LR-	TP	TN	FP	FN	Sen	Spec	Acc	1/LR+	LR-	TP	TN	FP	FN	Sen	Spec	Acc	1/LR+	LR-
'most'	4	4	1M/2F	3	20	0	1	0.75	1	0.95	0	0.25	3	19	1	1	0.75	0.95	0.91	0.12	0.27	0	20	0	4	0	1	0.83	NaN	1
'today'	3	5	2M/1F	5	30	3	0	1	0.90	0.92	0.09	0	4	31	2	1	0.8	0.93	0.92	0	0.6	0	33	0	5	0	1	0.86	NaN	1
'fish'	2	4	1M/1F	4	27	4	0	1	0.87	0.88	0.12	0	3	27	4	1	0.75	0.87	0.85	0.18	0.28	1	31	0	3	0.25	1	0.91	0.18	0.8
'again'	3	4	1M/1F	4	38	2	0	1	0.95	0.95	0.05	0	4	32	8	0	1	0.80	0.81	0.16	0.28	2	40	0	2	0.5	1	0.95	0	0.7
قُل	5	5	3M/2F	5	48	11	0	1	0.81	0.82	0.18	0	3	50	9	2	0.6	0.84	0.82	0.25	0.47	1	58	1	4	0.2	0.9	0.92	0.08	0.8
'dog'	1	4	1M/1F	4	27	2	0	1	0.93	0.93	0.06	0	1	26	3	3	0.25	0.89	0.81	0	0.75	1	29	0	3	0.25	1	0.9	NaN	1
'john'	3	5	2M/1F	5	32	5	0	1	0.86	0.88	0.13	0	4	30	7	1	0.8	0.81	0.8	0.16	0.23	0	37	0	5	0	1	0.8	NaN	1
'wood'	3	3	1M/2F	3	24	0	0	1	1	1	0	0	3	22	2	0	1	0.91	0.92	0.2	0.71	0	24	0	3	0	1	0.88	NaN	1
'collect'	3	3	1M/2F	3	27	1	0	1	0.96	0.96	0.03	0	2	27	1	1	0.6	0.96	0.93	0	0.33	0	28	0	3	0	1	0.9	NaN	1
'found'	2	4	2M/2F	4	40	1	0	1	0.97	0.97	0.02	0	4	35	6	0	1	0.85	0.86	0.1	0.27	3	41	0	1	0.75	1	0.97	NaN	1
'enough'	2	4	2M/2F	4	33	1	0	1	0.97	0.97	0.02	0	4	33	1	0	1	0.97	0.97	0.8	1.1	3	34	0	1	0.75	1	0.97	NaN	1
'cap'	3	4	2M/1F	3	18	1	1	0.75	0.94	0.91	0.07	0.26	3	19	0	1	0.75	1	0.95	0.7	0.91	0	19	0	4	0	1	0.82	NaN	1
اَلنّاس	2	5	2M	4	22	0	1	0.8	1	0.96	0	0.2	4	21	1	1	0.8	0.95	0.92	0.05	0.21	0	22	0	5	0	1	0.81	NaN	1
'bed'	1	4	2M/1F	4	36	5	0	1	0.87	0.88	0.12	0	4	37	4	0	1	0.90	0.91	0.05	0.26	0	41	0	4	0	1	0.91	NaN	1
'throw'	1	4	1M/2F	4	41	2	0	1	0.95	0.95	0.04	0	4	40	3	0	1	0.93	0.93	0	0.5	0	43	0	4	0	1	0.91	NaN	1
'tim'	3	4	2M/1F	4	45	11	0	1	0.80	0.81	0.19	0	4	47	9	0	1	0.83	0.85	0.25	0.57	4	56	0	0	1	1	1	NaN	1
'thought'	4	4	1M/1F	3	44	4	1	0.75	0.91	0.90	0.11	0.27	2	41	7	2	0.5	0.85	0.82	0.19	0.78	1	48	0	3	0.25	1	0.94	NaN	1
'dog'	3	4	1M/1F	4	43	4	0	1	0.91	0.92	0.08	0	4	42	5	0	1	0.89	0.9	0.26	0.8	2	47	0	2	0.5	1	0.96	NaN	1

‘decline’	3	4	1M/2F	4	31	5	0	1	0.86	0.87	0.13	0	4	30	6	0	1	0.83	0.85	0	0	1	36	0	3	0.25	1	0.92	NaN	1
‘said’	2	5	1M/2F	5	50	17	0	1	0.74	0.76	0.25	0	3	58	9	2	0.6	0.86	0.84	0.13	0.43	2	67	0	3	0.4	1	0.95	NaN	1
آپ	3	4	2M/1F	4	26	0	0	1	1	1	0	0	3	22	4	1	0.75	0.84	0.83	0.21	0.29	1	26	0	3	0.25	1	0.9	0	0.7
‘fish’	2	5	2M/2F	5	34	5	0	1	0.87	0.88	0.12	0	3	37	2	2	0.6	0.94	0.9	0.13	0.22	1	39	0	4	0.2	1	0.9	0	0.8
‘albert’	3	4	2M/1F	4	21	0	0	1	1	1	0	0	4	21	0	0	1	1	1	0.11	0.27	1	21	0	3	0.25	1	0.88	NaN	1
‘threw’	4	4	1M/2F	4	43	11	0	1	0.79	0.81	0.20	0	4	45	9	0	1	0.83	0.84	0	0.5	2	54	0	2	0.5	1	0.96	NaN	1
‘spect’	2	4	1M/1F	4	28	0	0	1	1	1	0	0	4	26	2	0	1	0.92	0.93	0.4	0.62	1	28	0	3	0.25	1	0.9	NaN	1
‘collect’	3	4	1M/1F	4	25	0	0	1	1	1	0	0	3	25	0	1	0.75	1	0.96	0.05	0.26	0	25	0	4	0	1	0.86	NaN	1
‘pilled’	3	4	1M/1F	4	33	5	0	1	0.86	0.88	0.13	0	4	28	10	0	1	0.73	0.76	0.2	0.55	2	38	0	2	0.5	1	0.95	NaN	1
‘please’	3	4	2M/2F	4	34	1	0	1	0.97	0.97	0.02	0	4	33	2	0	1	0.94	0.94	0.08	0	2	35	0	2	0.5	1	0.94	NaN	1
‘main Road’	4	4	2M/2F	4	12	0	0	1	1	1	0	0	3	12	0	1	0.75	1	0.93	0.15	0	1	12	0	3	0.25	1	0.81	NaN	1
دھماکہ	4	5	3M/1F	5	22	0	0	1	1	1	0	0	5	22	0	0	1	1	1	0	0	1	22	0	4	0.20	1	0.85	0	0.8
اَللہ	5	5	4M/1F	4	13	0	1	0.80	1	0.94	0	0.2	4	13	0	1	0.80	1	0.94	0	0.2	0	13	0	5	0	1	0.72	NaN	1
‘scare Me’	2	4	2M/1F	4	18	2	0	1	0.90	0.91	0.1	0	3	18	2	1	0.75	0.9	0.87	0.12	0	1	20	0	3	0.25	1	0.87	NaN	1
‘port’	4	4	2M/2F	4	27	0	0	1	1	1	0	0	4	24	3	0	1	0.88	0.9	0.37	0.61	1	27	0	3	0.25	1	0.9	NaN	1
‘quickly’	3	4	1M/1F	4	19	2	0	1	0.90	0.92	0.09	0	3	20	1	1	0.75	0.95	0.92	NaN	1	0	21	0	4	0	1	0.84	NaN	1
‘short’	3	3	1M/1F	3	38	4	0	1	0.90	0.91	0.09	0	3	35	7	0	1	0.83	0.84	0.45	0.78	0	42	0	3	0	1	0.93	NaN	1
Average of Sensitivity, Specificity, Accuracy, 1/LR+, LR-:							**0.97**	**0.93**	**0.92**	**0.06**	**0.03**						**0.83**	**0.90**	**0.88**	**0.17**	**0.43**					**0.24**	**0.99**	**0.89**	**NaN**	**0.96**

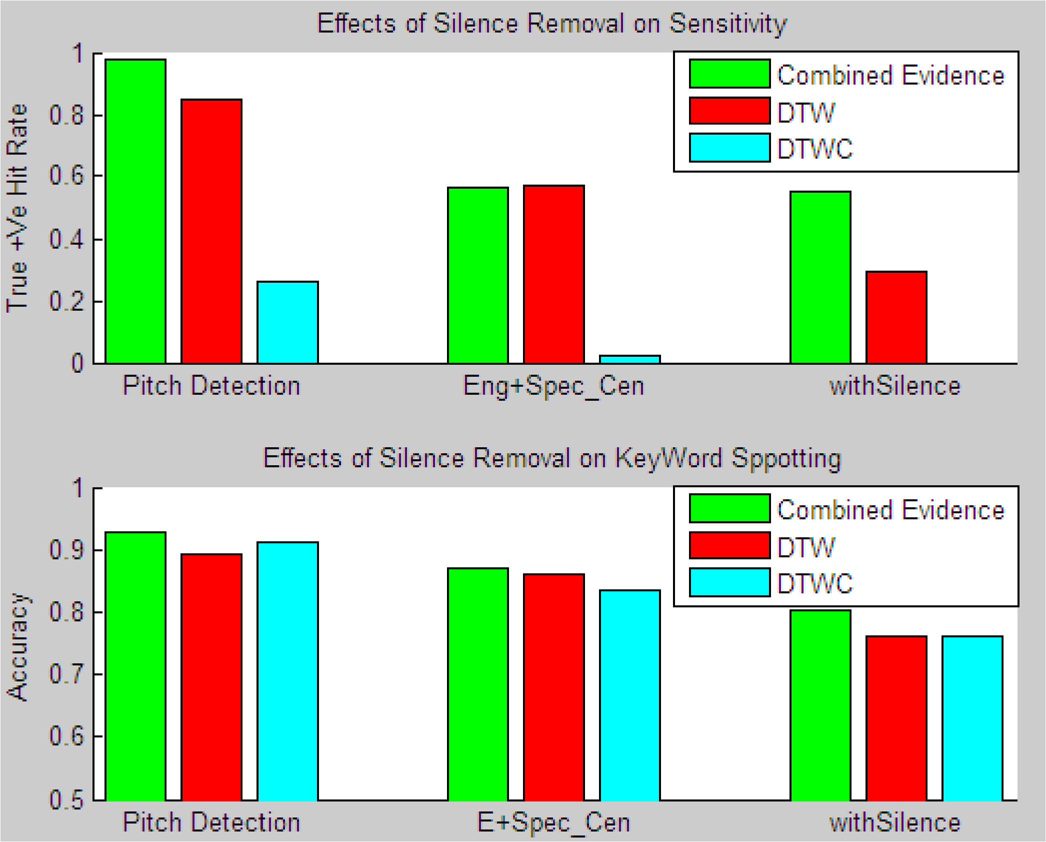


Figure 6-10: Performance Analysis for Different Silence Removal Approaches

Above figure demonstrates a good indication of silence removal impacts on different keyword spotting approaches. It is clear that the sensitivity is increased to 97.5% with the proposed pitch detection based silence removal (Khan and Holton 2015) as compared to 56% with the most commonly used energy and spectral centroid based approach (Sahoo and Patra 2014), (Sharma and Rajpoot 2013). It can also be observed that the traditional silence removal approaches don't effect the proposed CEv based approach merely but also reduce the performance of other approaches which validates the generalization of the newly introduced silence removal approach.

Table 6-2: Impact of Silence Removal Techniques and Performance Analysis of KeyWord Spotting Approaches in Terms of Execution Time and Mean Square Error

		Combined Evidence	DTW	DTW Constrained	Mean-MFCC	Wavelets
Performance withou Silence Removal	Accuracy	0.8044	0.7616	0.7624	0.8102	0.8265
	True +Ve Rate	0.5544	0.2928	0	0.5739	0.7806
Performance with Energy & Spectral Centroid Based Silence	Accuracy	0.8687	0.8597	0.8340	0.8706	0.8608
	True +Ve Rate	0.5656	0.5728	0.0233	0.5600	0.7617
Performance with Energy and Pitch Detection Based Silence	Accuracy	0.9266	0.7616	0.9110	0.9301	0.8695
	True +Ve Rate	0.9750	0.2928	0.2617	0.9572	0.9322
Type I Error	µ	0.0105	0.0105	6.8871e-05	0.0090	0.0258
	σ	0.0148	0.0110	3.7722e-04	0.0118	0.0260
Type II Error	µ	0.0063	0.2596	0.9588	0.0113	0.0300
	σ	0.0191	0.2762	0.1262	0.0274	0.1042
Avg. Ex. Time Per Test Case (Seconds)		1.3483	1.4445	0.0826	0.0672	1.2604

Type I and Type II errors indicate the recogniser failure related to FP and FN respectively. These metrics have been represented in a number of ways in the related area including mean square error and absolute errors as the most common metrics. Table 6-2 demonstrates 'µ' (mean) and 'σ' (standard deviation) for both types of error for five different approaches while using the same dataset. As discussed before, in keyword spotting related tasks, Type II error may have more importance as compared to Type I error because of the more emphasis on spotting a keyword. However, it may vary with respect to application area. It can be observe from Table 6.2 that the 'µ' and 'σ' for Type II error are negligible (i.e. 0.006 and 0.019 respectively) in case of CEv based keyword spotting as compared to DTW approach (i.e. 0.25 and 0.27 respectively) which indicates the robustness of the proposed approach. In addition, it can also be observed that the individual errors for MFCCs and wavelet based approaches are higher than the CEv approach that proves the significance of the deployment of DST for keyword spotting task.

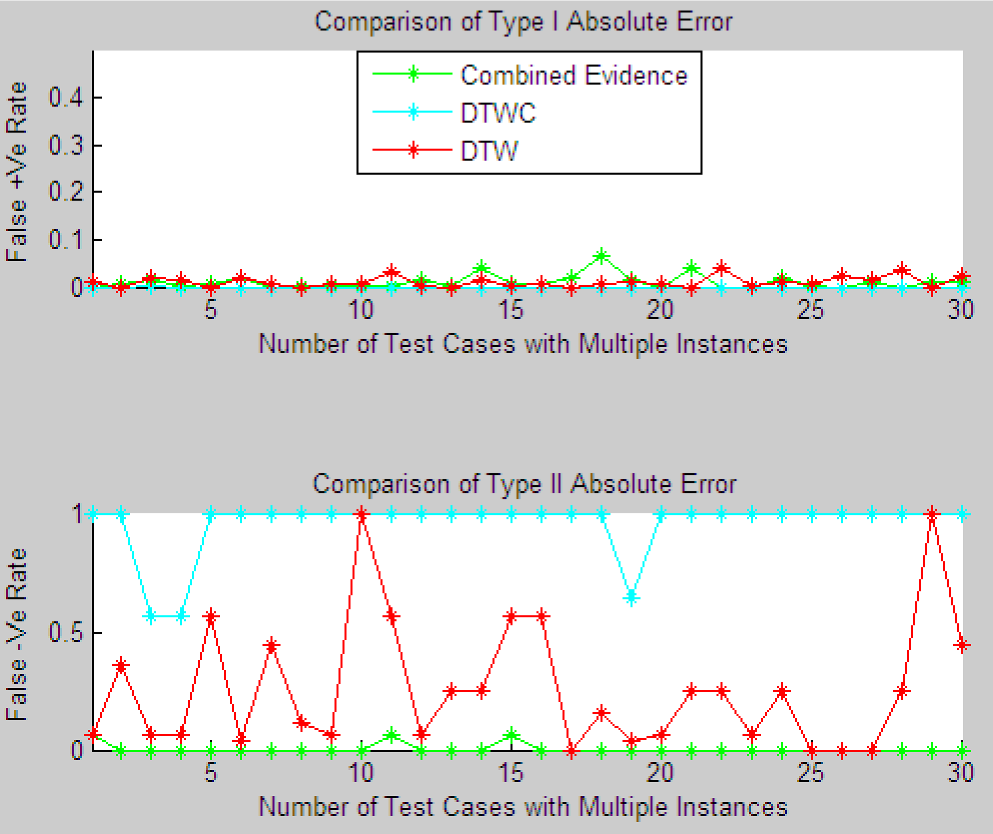


Figure 6-11: Absolute Error for Multiple Keyword Spotting Approaches

Despite of the keyword spotting performance of the aforementioned approaches, it is also important to analyse the computation time. This may help to analyse the workload increment due to additional processing of DST application. Figure 6-12 provides an indication of execution time for all aforementioned approaches. It is quite clear that the best execution time is acheievd by MFFC based approach that is introduced in the current research study. This is because of Euclidean distance deployemnt for mean values of MFCCs features in the current research rather than DTW which increases the search space (Cheng-Tao et al. 2014) as it has been used in the literature. Although, the minimum execution time (i.e. 0.06 sec) is achieved by MFCC based approach, yet CEv approach with higher excution time (i.e. 0.3 sec) would be preferred because of its superiority in terms of keywords detection rate which is the desired objective.

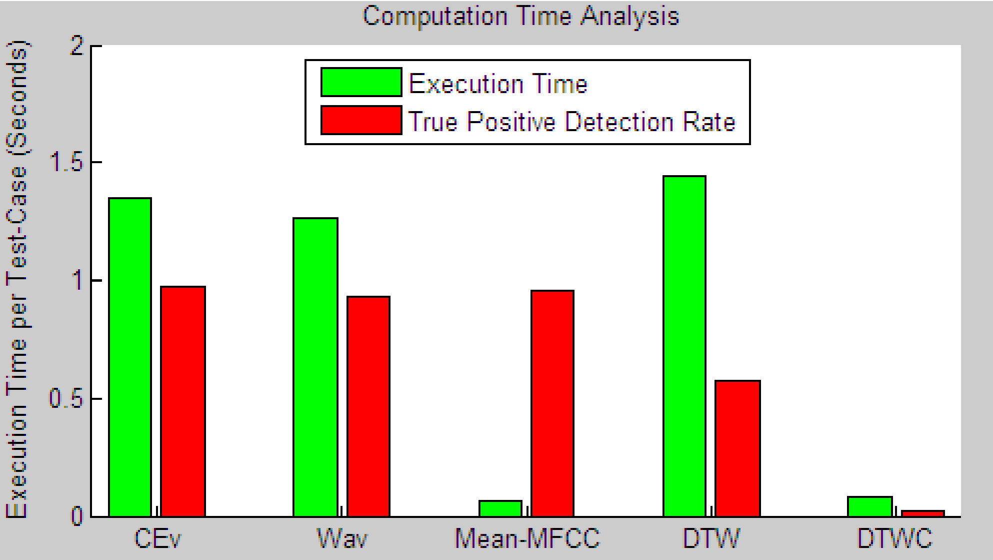
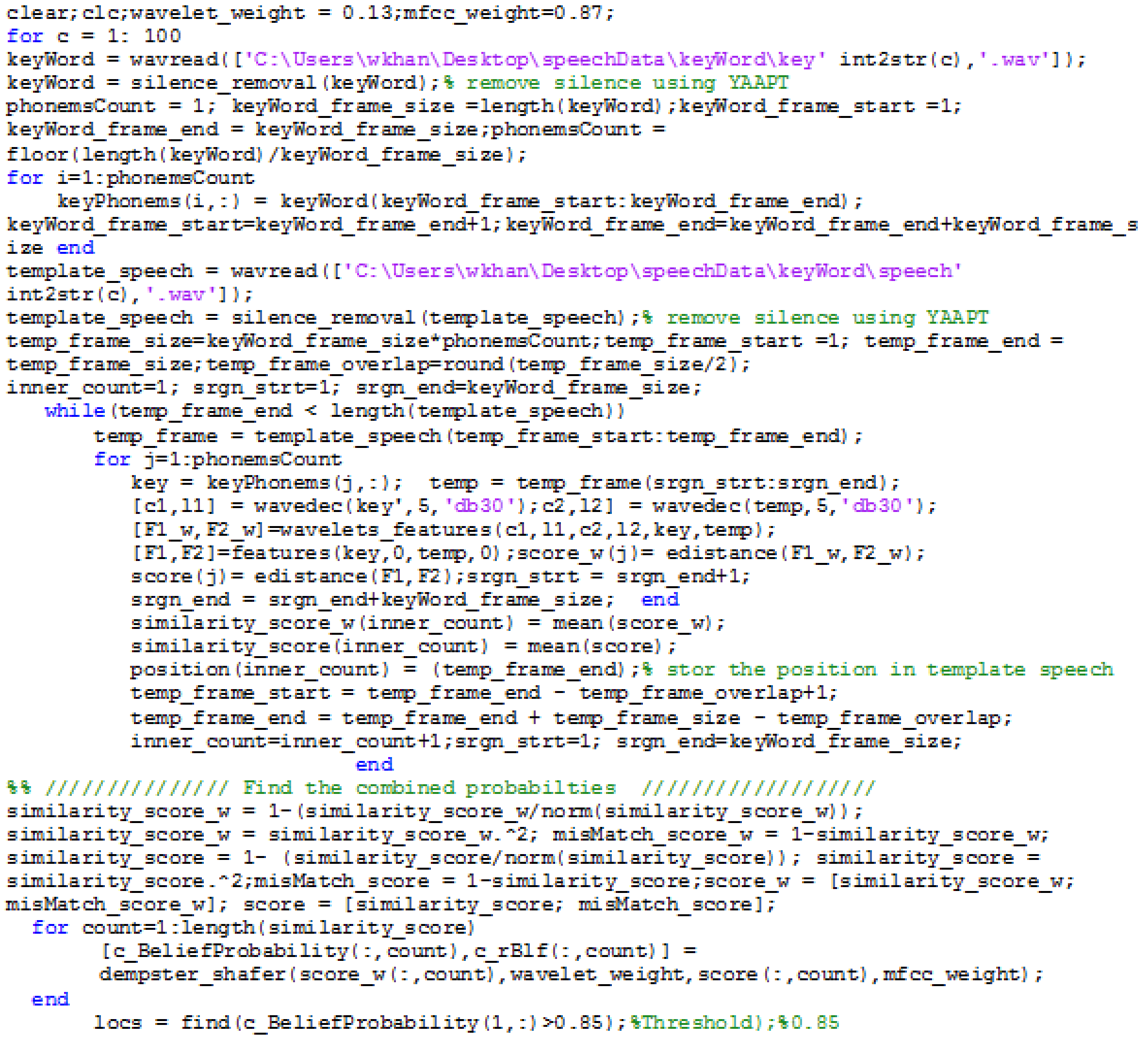


Figure 6-12: Trade-Off between Keyword Detection Rate and Computational Cost

The impact of DTW constraints in terms of pruning as discussed in Chapter 3, Section 3.4 can be observe in Figure 6-12. It is quiet clear that the the execution time is dramatically decreases from 1.4 seconds to 0.1 seconds by applying constaints on the traditional DTW. However, it sacrifies a significant amount of keyword detection rate (50%) that fails the achievement of the primary objective.



```
clear;clc;wavelet_weight = 0.13;mfcc_weight=0.87;
for c = 1: 100
keyWord = wavread(['C:\Users\wkhan\Desktop\speechData\keyWord\key' int2str(c),'.wav']);
keyWord = silence_removal(keyWord);% remove silence using YAAPT
phonemsCount = 1; keyWord_frame_size =length(keyWord);keyWord_frame_start =1;
keyWord_frame_end = keyWord_frame_size;phonemsCount =
floor(length(keyWord)/keyWord_frame_size);
for i=1:phonemsCount
    keyPhonems(i,:) = keyWord(keyWord_frame_start:keyWord_frame_end);
keyWord_frame_start=keyWord_frame_end+1;keyWord_frame_end=keyWord_frame_end+keyWord_frame_s
ize end
template_speech = wavread(['C:\Users\wkhan\Desktop\speechData\keyWord\speech'
int2str(c),'.wav']);
template_speech = silence_removal(template_speech);% remove silence using YAAPT
temp_frame_size=keyWord_frame_size*phonemsCount;temp_frame_start =1; temp_frame_end =
temp_frame_size;temp_frame_overlap=round(temp_frame_size/2);
inner_count=1; srgn_strt=1; srgn_end=keyWord_frame_size;
   while(temp_frame_end < length(template_speech))
       temp_frame = template_speech(temp_frame_start:temp_frame_end);
       for j=1:phonemsCount
          key = keyPhonems(j,:);  temp = temp_frame(srgn_strt:srgn_end);
          [c1,l1] = wavedec(key',5,'db30');c2,l2] = wavedec(temp,5,'db30');
          [F1_w,F2_w]=wavelets_features(c1,l1,c2,l2,key,temp);
          [F1,F2]=features(key,0,temp,0);score_w(j)= edistance(F1_w,F2_w);
          score(j)= edistance(F1,F2);srgn_strt = srgn_end+1;
          srgn_end = srgn_end+keyWord_frame_size;  end
          similarity_score_w(inner_count) = mean(score_w);
          similarity_score(inner_count) = mean(score);
          position(inner_count) = (temp_frame_end);% stor the position in template speech
          temp_frame_start = temp_frame_end - temp_frame_overlap+1;
          temp_frame_end = temp_frame_end + temp_frame_size - temp_frame_overlap;
          inner_count=inner_count+1;srgn_strt=1; srgn_end=keyWord_frame_size;
                            end
%% /////////////// Find the combined probabilties  ///////////////////
similarity_score_w = 1-(similarity_score_w/norm(similarity_score_w));
similarity_score_w = similarity_score_w.^2; misMatch_score_w = 1-similarity_score_w;
similarity_score = 1- (similarity_score/norm(similarity_score)); similarity_score =
similarity_score.^2;misMatch_score = 1-similarity_score;score_w = [similarity_score_w;
misMatch_score_w]; score = [similarity_score; misMatch_score];
  for count=1:length(similarity_score)
        [c_BeliefProbability(:,count),c_rBlf(:,count)] =
        dempster_shafer(score_w(:,count),wavelet_weight,score(:,count),mfcc_weight);
  end
       locs = find(c_BeliefProbability(1,:)>0.85);%Threshold);%0.85
```

Figure 6-13: Matlab Script for Keyword Spotting

Figure 6-13 presents a Matlab script for the proposed keyword spotting. Complete scripts corresponding to each component of the flowchart in Figure 6-2 are presented in Appendix D.

6.4. Word Matching Based on Posterior Probability Measure

Speech signal may consist of background noise that reduces the similarity matching performance. Differentiation between background noise and speech signal may be one possible solution to avoid the mismatch or at least improve the performance. To predict the target word in template speech, it may be helpful to search an estimated segment rather than the entire signal. The area searched for a match is called a *search region* (S). The matching task proceeds along the speech signal by overlapped frames. The most similar matched frame is called *target region*. During localisation of the target keyword, it is certain that the background noise components of '*S'* are mixed with the template speech. When speech quality is poor, these components may produce interference in localisation of the true target and lead to biased localisation or misidentification. One way to prevent misidentification of target speech is the differentiation between keyword features and background noise features.

Usually '*S'* is larger than keyword which means there exist more background noise components than the keyword components. Consequently, feature vectors of the search region can provide clues to understanding the statistical characteristics of the background noise components. A larger value of 'S_u' indicates feature '*u'* is more likely to be a feature of background noise, whereas for a sample feature of keyword, 'S_u' is more likely to be small. This consideration leads to the introduction of '1/S_u' as a rectifying weight for the reduction of background noise influence on actual speech. This idea is introduced very first time by (Fing et al. 2008) as a PPM for image matching that leads to adaptive scaling. The aforementioned consideration is utilized by using the cross-correlation as a prototype to get the new similarity measure, PPM. Mathematically;

$$\phi(A,B) = \frac{1}{m}\sum_{u=1}^{m_u}\frac{A_u B_u}{S_u} \quad 6\text{-}17$$

Where 'S_u', 'A_u', and 'B_u' represent the u^{th} element of un-normalised feature vectors 'S', 'A', and 'B' respectively; 'A' and 'B' are the feature vectors of the keyword and the template speech respectively; 'm' is a normalisation constant representing the number of components in keyword. Consideration of the background noise components is the major advantage of PPM over Bhattacharyya and other existing similarity measure techniques. The similarity is measured by correlation between template and keyword feature components.

The aim is to search for a keyword/target model (TM) in a short speech phrase consisting 4 words at maximum including 'TM' at least once. The template speech utterance is divided into small segments with size equal to 'TM'. The overlapped segments that are matched against the 'TM' are named target candidates (TC). Both, 'TM' and the 'TC' are characterised by a same size features vector (i.e. MFCCs, section 2.2.3). Thus, we have:

$$TM : \mathrm{A} = \{A_u\}_{u=1.....\mathrm{m}_u} \tag{6-18}$$

$$TC : \mathrm{B} = \{B_u\}_{u=1.....\mathrm{m}_u} \tag{6-19}$$

Where 'A' and 'B' represent the 'TM' feature vector and 'TC' feature vector respectively and 'm_u' is the dimension of frequency vectors. A block diagram is presented below that shows the sequential processing for PPM based keyword identification.

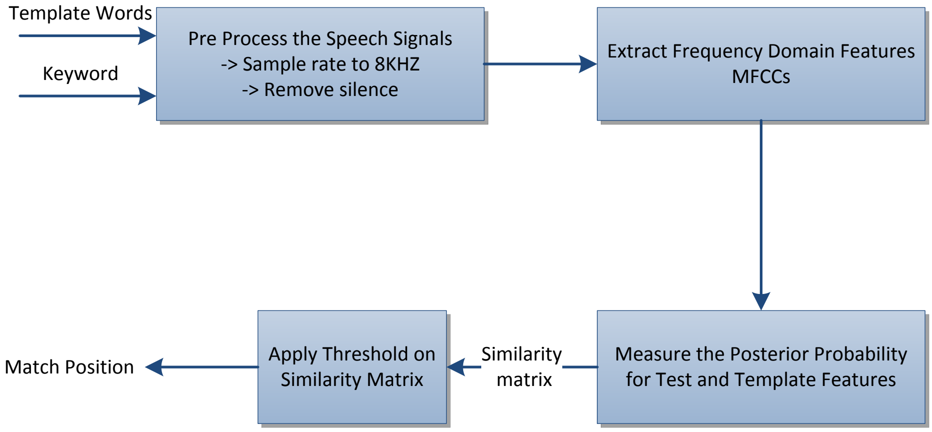


Figure 6-14: Block Diagram for PPM Based KeyWord Matching

In the pre-treatment component, template and input speech utterances are resampled to 8 KHz. After resampling; the silence segments are removed from the speech signal using the energy and spectral centroid method (Chapter 2, section 2.2). In next stage, the enhanced signal from previous step is divided into small frames with size equal to ‘TM’. The framed data and ‘TM’ are then converted into feature vectors which are used for further process of similarity match. Generally, human voice contains important information such as gender, emotion, and identity of speaker that can be categorised in different classes. Extracted features provide better results when they do not loose class related information. In this approach, MFCC feature vectors are used to represent the speech signals. Detailed implementation of MFCCs is presented in Chapter 2, section 2.2.3.

The PPM is then applied to the feature vectors of ‘TM’ and ‘TC’ to find out the similarity score for ‘TM’. Based on this score, position of the ‘TM’ in reference speech is measured. This is achieved by calculating the start and end positions of ‘TM’ in ‘RM’. The similarity measure results of overlapped windows are saved into one dimensional matrix called similarity matrix (SM). The start and end positions of ‘TM’ can be localised in reference speech as:

$$Y_1 = \frac{Y * X_1}{X}, \quad Y_2 = \frac{Y * X_2}{X} \qquad \text{6-20}$$

where ‘X’ represents the number of samples in speech phrase, ‘X_1’ is start position of ‘TM’ in speech sentence, ‘X_2’ is the end position of ‘TM’ in speech sentence, ‘Y’ represents the number of indexes in ‘SM’, ‘Y_1’ and ‘Y_2’ are the resulting start and end indexes within the ‘SM’ respectively that represent the most similar segment for ‘TM’ in the speech sentence. The word recognition and localisation result is perfect if the similarity values from ‘Y_1’ to ‘Y_2’ are higher than rest of ‘SM’ indexes values as shown in performance analysis below.

6.4.1. Performance Analysis of PPM

The inability of existing speech similarity measures to efficiently differentiate between target frequency components and background noise components downgrades their performance. Existing techniques work well with word recognition in the absence of background noise but produce bad performance in the presence of background noise that is not the case for PPM which is capable of differentiating the target components from background noise components while performing the match operation. This feature enables the PPM to provide comparatively sharper match results even in the presence of background noise. For the first time the PPM is deployed as a similarity measure for the speech signal matching (Khan et al. 2012). The use of PPM for speech signal produced improved results not only for isolated word similarity measure but also for the word localisation in a short continuous speech phrase.

In the experimental results presented in Table 6-3, a public crowd noise is added to speech data as background noise. Results are achieved and compared for different level of intensity in background noise using Matlab R2009a. Table 6-3 demonstrate the overall performance comparison of PPM with Bhattacharya and cross correlation approaches. The Accuracy metric is measured using the confusion matrix as described earlier in this chapter. A small dataset with 60 recordings of 3 keywords is used. It can be observe that the performance of PPM is better than other two approaches not only in in normal speech but also, it is better in case of noise addition.

Table 6-3: Performance Evaluation of PPM and Existing Similarity Measures For Keyword Identification (Khan et al. 2012)

Keyword	No. of Template Speech Phrases	Accuracy (%) Without Noise			Accuracy (%) with Background Noise of Amplitude [-0.2 0.2] i.e. SNR is 2:1		
		PPM	Bhata	Cross Correlation	PPM	Bhata	Cross Correlation
Hello	25	98.1	90	70.3	98	45	43
computer	20	90	60	67.8	90	30.1	27.2
slow	15	100	89	85	100	52	49

A test case is presented in Figure 6-15 and Figure 6-16 for the performance of PPM, Bhattacharyya and cross correlation in noise reduced and noise added speech respectively. The keyword '*hello'* which represents the 'TM' is matched against the continuous speech phrase with the contents of '*One Two Hello One Two'*. These figures demonstrate the difference between the performances of the PPM with Bhattacharyya, and cross correlation techniques with sharp peaks. It is evident that the performance of existing techniques shows dramatic change in the peaks when background noise is added (Figure 6-16). Compared to these techniques, the PPM remained stable and indicated accurate localisation of keyword.

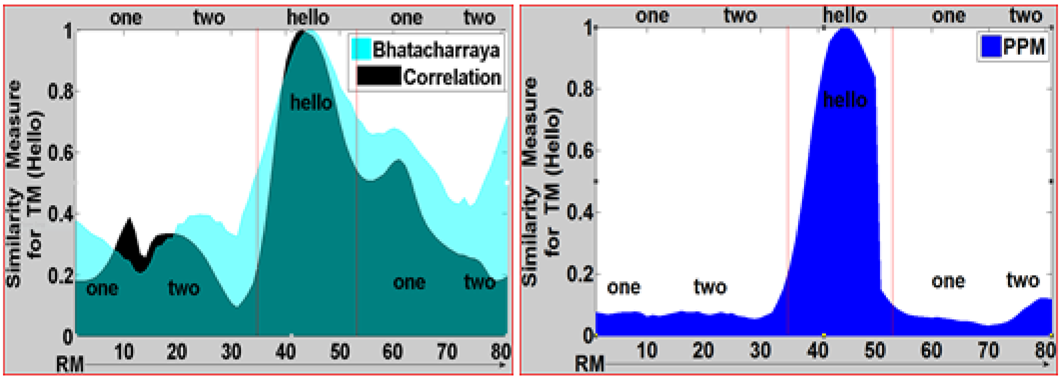


Figure 6-15: Keyword Localization with Noise Reduced Speech (Khan et al. 2012)

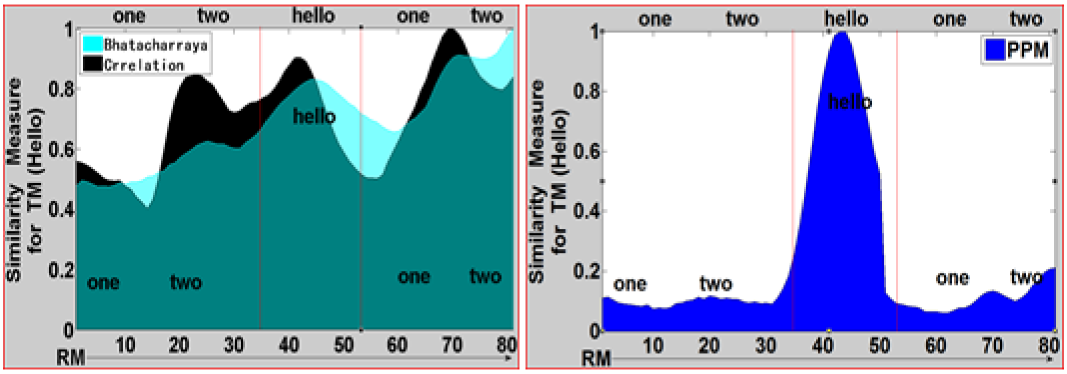


Figure 6-16: Keyword Localization with Noise Added Speech (Khan et al. 2012)

Figure 6-17 shows performance of the PPM, Bhattacharyya and cross correlation approaches by adding different intensity levels of background noise in the speech sentence. The noise added in speech signal is public crowed noise. It is observed in Figure 6-17 that the Bhattacharyya and cross correlation techniques produce poor results with mixture of very low intensity background noise whereas PPM can recognise the target word until speech to background noise ratio is less than 1:4.

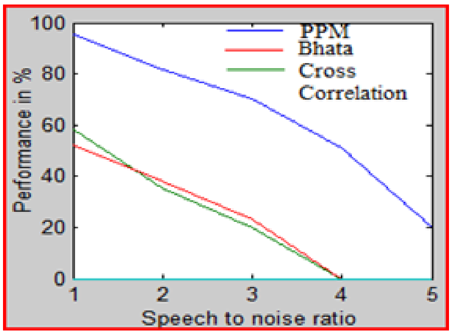


Figure 6-17: Performance Evaluation with Noise Added Speech (Khan et al. 2012)

6.5. Vector Addition Based Similarity Measure

Another compact but interesting contribution related to similarity metrics is introduced in the proposed research work that is based on vector addition (Khan et al. 2014). The aim of this study is to compare different similarity metrics in terms of their mathematical formulation and prove the concept using isolated word identification. Euclidean distance, vector cosine angle distance, Bhattacharyya coefficients, and PPM are used for the analysis and comparison purpose. The results show that the proposed similarity measure provides a satisfactory performance in terms of keyword matching when compared to aforementioned approaches (Khan et al. 2014). This metric uses the concept of vector addition to measure the similarity measurement between two vectors of n-dimensions. The resultant vector provides the similarity calculation between two vectors which is quite similar to the cosine of the angle between the two vectors. As compared to aforementioned algorithms, the computational cost of this approach is very low due to the arithmetic operator's simplicity. Mathematically;

$$\begin{aligned} RS_{(AB)} &= \frac{A}{\|A\|} + \frac{B}{\|B\|} \\ &= \left(a_1 + b_1\right) + \left(a_2 + b_2\right) + \ldots.. + \left(a_n + b_n\right) \end{aligned} \qquad 6\text{-}21$$

Where, ||A|| and ||B|| represents the 'norm' of the vectors 'A' and 'B' respectively, 'RS' is the resultant similarity measure, a_{i} and b_{i} are the i^{th} components of vector 'A' and 'B' respectively.

6.5.1. Theoretical Justification and Performance Analysis of RS

Cosine similarity provides the results that are based on the angular difference between the vectors. If we compare the RS with cosine similarity between unit vectors, it is clear that there is a direct relation between RS and CAS that implies mathematical justification of RS. Taking square of Equation 6-21, we get:

$$(|\hat{A} + \hat{B}|)^2 = |\hat{A}|^2 + |\hat{B}|^2 + 2|\hat{A}||\hat{B}| \qquad 6\text{-}22$$

In case of unit vectors (normalised), the above equation can be represented as:

$$RS^2 = 1 + 1 + 2\cos\theta$$

$$RS^2 = 2(1 + \cos\theta) \qquad 6\text{-}23$$

Above equation provides the mathematical justification of a relationship between CAS and RS. A number of aspects can be considered for comparison of ‘RS’ with vectors cosine angle and Euclidean distance. Firstly, ‘RS’ provides the degree of similarity measure rather than difference or dis-similarity as in Euclidean distance. Secondly, it resolves the Euclidean distance normalization issue. In Figure 6-18, ‘RS’ indicates a higher similarity between ‘AC’ in terms of direction as compare to ‘AB’ (i.e. RAC > RAB) which will not be in the case of Euclidean distance due to un-normalized vector lengths. Finally, it provides almost same results as cosine similarity because of its linear relationship to CAS. In addition, the computational cost is lower for ‘RS’ as compared to CAS because it alters the multiplication operator with addition of unit vectors.

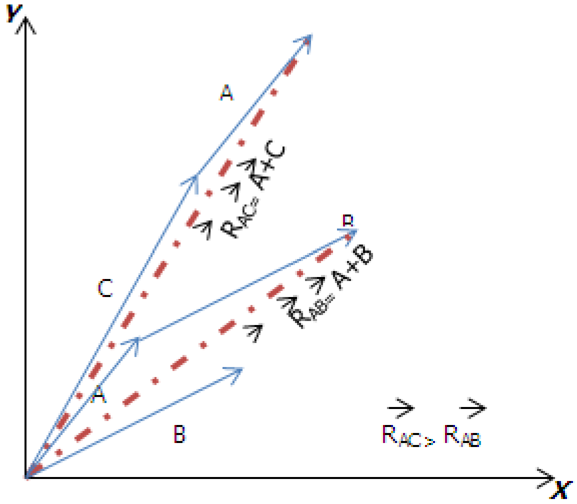

Y
A
$\vec{R}_{AC}= \vec{A}+\vec{C}$
B
C
$\vec{R}_{AB}= \vec{A}+\vec{B}$
A
B
$\vec{R}_{AC} > \vec{R}_{AB}$
X

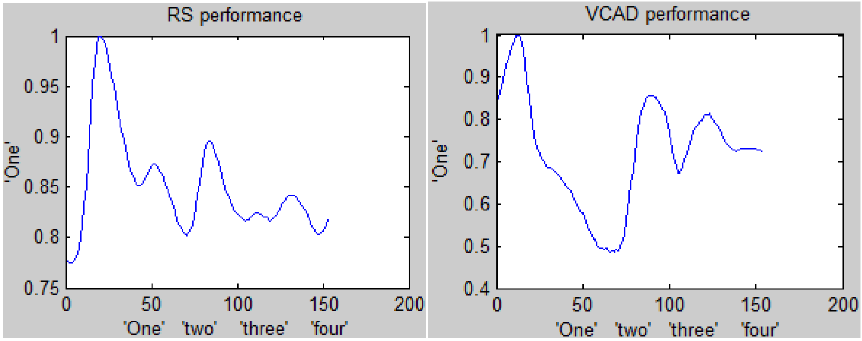
Figure 6-18: Vector Addition Based Similarity Measurement

The similarity performance of RS is compared to most commonly used similarity measures including CAS, ED, and Bhattacharyya. The experiments are conducted on a small dataset consists of isolated English words, acquired from a number of speakers from different age groups and gender. Matlab R2013a is used for simulations and experimental results. Table 6-4 shows the overall performance comparison of RS with other similarity measure in terms of matching for different length of template speech signals recorded by variety of speakers with different age and gender. Speech signal is enhanced before the application of spectral analysis and similarity measure. This enhancement improves the speech signal quality in terms of noise filtration and silence removal and sample rate conversion.

Table 6-4: Performance Evaluation of RS, CAS, ED, and Bhattacharya (Khan et al. 2014)

Target Keyword	No. of Template Speech	Spotting Accuracy (%)			
		RS	Cosine	PPM	Bhattacharya
‘Hello’	35	74	75	76	69
‘Computer’	63	71	70	72	70.5
‘Five’	77	69	71	74	68

It can be observed in Table 6-4 that the similarity matching performance of RS is approximately equal to that of CAS because of its linear relation and slightly different from others. Theoretically, it performs better than all of these similarity measures in terms of time complexity due to the usage of an addition operator instead of multiplication while not degrading the similarity matching performance. A test case performance of RS, CAS, Bhattacharya, and Euclidean distance for a short speech phrase is presented in Figure 6-19. It can be analysed that the performance result of all similarity measure is almost equal except Bhattacharyya. These results validate the RS as a new similarity metrics that is linearly related to cosine angle based similarity measure.



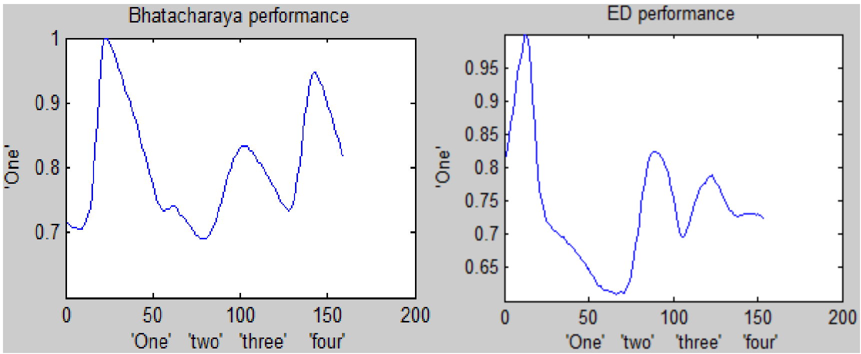


Figure 6-19: Similarity Measure of RS, CAS, ED and Bhattacharya (Khan et al. 2014)

6.6. Summary

In this chapter, a comprehensive overview of the research contribution towards the keyword spotting in continuous speech is presented. Very first time, Dempster's theory of mass combination is deployed into the area of speech processing and keyword spotting. The idea is to combine beliefs of more than one resource (i.e. distance metrics/features etc.) in the form of similarity score. This empowers the final decision for whether the key-word being matched to a specific frame in the template speech or not? A detailed flowchart for sequential processing of keyword spotting is presented. Performance of the newly introduced keyword spotting approach is evaluated using the gold standards metrics for a binary classifier (i.e. match/mismatch). Another contribution in the related area is also addressed in the form of a posterior probability measure. A PPM is used for the first time in the current research study to localise an isolated keyword in a short speech phrase (e.g. 4 to 5 words at most) in presence of the background noise. The advantages of PPM are discussed over the related distance metrics and experimental results are discussed. Finally, a vector addition based similarity measure is introduced that provides an alternative to cosine similarity measure with less computational cost.

7. CONCLUSION AND FUTURE WORK

The hypothesis underlying the principal research question as motivated for this research study is based on speaker dependent time warped continuous speech tracking and similarity measurement. A comprehensive study of the latest advances is conducted, as reported by leading researchers in the field of speech signal matching approaches. To achieve the research aim defined in first research question in terms of TWCST approach, an experimental setup comprising speech enhancement, adaptive framing, feature extraction, dynamic state model, feedback system, and similarity metrics is build up to conduct the experiments and validate the proposed approaches that are presented in Chapter 4 and 5. An efficient keyword spotting approach is introduced in Chapter 6 that comprises a number of methodologies leading to the achievement of the secondary research aim defined in research question 2. The evaluation results using the experimental system with the various speech corpuses for benchmarking have established the validity of the underlying hypothesis in a replicable fashion, thus providing a positive response to the primary research questions. The aforementioned research contributions were made following the research methodology defined in Section 1.3. The research findings have proved the speech tracking and keyword spotting tasks achievable with a performance exceeding the state-of-the-art techniques. As the TWCST and keyword spotting approaches comprise of multiple techniques related to different areas, research contributions are also presented in diverse areas corresponding to research objectives defined in Chapter 1, section 1.3 as follows:

Dynamic State Model: To achieve the research objective **A(e),** a dynamic state model is introduced for TWCST and similarity measurement approach that considers the test speech signal as a linearly moving object along the template speech signal as presented in Chapter 4, section 4.3. Both signals are analysed frame by frame (i.e. time interval) and the best matched position is identified for the current iteration. The estimated position is forwarded as an input parameter to a Kalman filter for further processing (Khan and Holton 2015), (Khan et al. 2014).

Kalman Filter and Position Estimation in Continuous Speech Tracking: The use of KF is novel for speech template matching and similarity measurement. To achieve the research objective addressed in **A(g),** the TWCST approach presented in Chapter 4, uses two position observations from a DSM and similarity measure algorithm and forward them to a KF along with the process and measurement noise covariance. The KF process these inputs and provides a best position estimate in the template signal corresponding to test speech frame at current time step. This position estimate is further processed by adaptive framing process to predict the new template frame size for next time step. The whole cycle runs recursively until the end of test or template speech signal and best estimated positions in the template speech relative to input speech frames are recorded along with the similarity scores (Khan and Holton 2015), (Khan et al. 2014).

Dynamic Time Warping and Frame Size Adaptation: To deal with the similarity matching of two time warped speech signal, DTW with numerous variations have been considered as state-of-the-art that produces accumulated minimum warped distance between two speech signals (Cassisi 2012), (Chan and Lee 2010), (Thambiratmann and Sridharan 2007), (Zhang and Glass 2011), (Carlin et al. 2011), (Zhang and Glass 2010), (Jansen et al. 2010), (Zhang et al. 2012). In relation to research objective **C**, a novel idea of frame size adaptation is introduced (Chapter 4, section 4.3) that changes the template speech frame size dynamically at each time step with respect to input speech. Initially; the template frame size is kept same as of test frame but it dynamically changes at each time step with respect to the Kalman filter's best position estimate for the current iteration. The iterative frame size adaptation implies that rather than finding the overall warped distance as in traditional DTW, the warped distance can be minimised recursively by dynamically adapting the template frame size relative to spoken utterance length (Khan and Holton 2015).

Dynamic Filtration of Speech Signal Based on Wavelet Decomposition: An effective approach is introduced in Chapter 5 using wavelet decomposition that is able to filter out unnecessary segments from speech signal dynamically at each time step. In Fourier transform, there is a trade-off between time and frequency resolution. Therefore, the entire frequency spectrum is needed to be mapped to obtain the similarity score between test and template speech frames. However, in case of wavelet decomposition, speech signal is presented in simultaneous time frequency view. The whole frequency band that doesn't exist in speech signal is ignored. Also, the corresponding time domain segment is also filtered out. The filtered form of spectrum is forwarded for the similarity score calculation. Statistical results are conducted on a variety of dataset that validated the proposed approach leading to the achievement of research objective addressed in **D** and **F** (Khan and Holton 2014).

Distance Metrics: To address the research objective **A(f)** and **E**, a detailed research is conducted in relation to distance metrics approaches. Euclidean distance and cosine similarity measure has been used extensively for similarity measure and probabilistic distribution. An innovative technique for image processing that was proposed by (Fing et al. 2008) known as posterior probability measure is deployed for the first time in proposed research work (Khan et al. 2012) for the speech signal matching as presented in Chapter 6, section 6.4. The PPM has a unique property of differentiating the background noise components from speech components when measuring the similarity between target and reference speech models. The separation of these features shared by both target and background models produced robust results in keyword localisation. It results more reliable and tolerant pattern match to varying model scale. Experiments proved that the PPM produces efficiency in keyword localisation whereas Bhattacharyya and other pattern matching techniques result in mismatch or bias due to interference of background components. In addition, a new distance metric is introduced (Chapter 6, section 6.5) based on vector addition (Khan et al. 2013).

Keyword Spotting Approach: Keyword spotting is an emerging research area that deals with the speech dynamics. Finding a keyword in a continuous speech signal is very useful in practical life. However, the existing approaches are not able to deal with speech dynamics to produce a reliable keyword spotter. To address the research objective **B**, an efficient keyword spotting approach is presented in Chapter 6. For the first time, Dempster’s theory of mass combination is deployed into the area of speech processing and keyword spotting. The idea is to combine beliefs from more than one information resources (i.e. distance metrics/features etc.) in the form of similarity scores. This empowers the final decision for whether the keyword being matched to a specific frame in the template speech or not? Performance of the newly introduced keyword spotting approach is evaluated using the gold standards metrics for binary classifier (i.e. match/mismatch) that proved the superiority of the proposed approach over the existing methodologies (Khan and Holton 2015).

7.1. Future Research Directions

Statistical results achieved and the experimental setup build and tested in attempting to validate the hypothesis posed in this PhD research study, serves as an enabler for future research on the following aspects;

- **Adaptive Feature Selection**

The proposed speech tracking approach can be enhanced using an adaptive feature selection approach. Feature selection, also known as variable selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that many redundant or irrelevant features exists in the data which provide no more information than the currently selected features. There are various feature selection methods that can be utilized. For example, greedy forward selection, correlation feature selection, entropy based feature selection, and genetic algorithms to find out the best feature set for corresponding speaker. Different feature set may produce different performance results depending upon speaker's linguistic features. Similarly, feature set may vary for different language phonemes to efficiently represent the acoustic features for that language model. In addition, a hybrid feature selection can also be used that combines the MFCC, LPC, and PLP etc., with time domain features.

- **Speaker Independence**

Speaker independence may be an interesting but a challenging task that can be considered as a future research direction of the proposed approach. Proposed approaches for TWCST and keyword spotting can be more useful in terms of their application areas in case of speaker independence. Traditional ASR systems produce the functionality of speaker independence with the use of transcribed data and training the system on statistical models. However, the proposed approaches are based on acoustic features that do not use the transcribed data. The solution may be constructed using the speaker normalization that can be achieved by normalizing some of the features of test speech speaker according to the template speech speaker. Pitch normalization may be a good example of such directive. Similarly, vocal tract normalisation is also a related active research area that can be considered to achieve the speaker independence functionality.

- **Combination of the Theory of Evidence and Kalman Filter**

The proposed research work for keyword spotting has validated the advantages of using the theory of evidence in speech signal matching area. In the future, it may also be combined with Kalman filter in TWCST approach to empower the matching decision. For test and template frame similarity measure, the position beliefs of a dynamic state model and feature based distance metric can be forwarded to combination of masses (Dempster-Shafer is one of the best approach for mass combination) and a combined belief of match/mismatch can be calculated (as discussed in Chapter 6). This belief along with a dynamic state model output and noise variances can be forwarded to Kalman filter to measure the desired position.

REFERENCES

Abad, A., Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., Bordel, G., (2013) 'On the Calibration and Fusion of Heterogeneous Spoken Term Detection Systems', *Conference of the International Speech Communication Association,* (ISCA) INTERSPEECH, Pp. 20-24, France, 25-29 August.

Allen, J. B. (1977) 'Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform', *IEEE._J._ASSP*, 25(3), Pp. 235-238.

Allen, J. B., Rabiner, L. (1977) 'A Unified Approach to Short Time Fourier Analysis and Synthesis'. *In. Proc. IEEE*. 65(11), Pp. 1558-1564. doi: 10.1109/PROC.1977.10770.

Anguera, X., Metze, F., Buzo, A., Szoke, I., Rodriguez-Fuentes, L. J. (2013) 'The spoken web search task', *in Proceedings of MediaEval*, *CEUR Workshop Proceedings*, Aachen, Germany, Pp. 1–2.

Anguera, X., Rodriguez-Fuentes, L. J., Szoke, I., Buzo, A., Metze, F. (2014) 'Query by example search on speech', *in Proceedings of MediaEval*, Pp. 1–2, Spain, October 16-17.

Arora, S. J., Singh, R. P. (2012) 'Automatic Speech Recognition: A Review', *International Journal of Computer Applications*, 60(9), Pp. 34-44.

Bahi, H & Benati, N. (2009) 'A New Keyword Spotting Approach'. *IEEE, International Conference on Mujltimedia Computing and Systems*, ICMCS, Pp. 77-80, Ouarzazate, 2-4 April.

Baker, J. M., Deng. L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., O'Shaughnessy, D. (2009) 'Research Developments and Directions in Speech Recognition and Understanding', *IEEE Signal Processing Mag*, Vol. 26, Pp. 75–80, May 2009.

Bob, L. (2009) '*Fundamentals of Real Time Spectrum Analysis'*, Tektronix Primer. Available at: http://materias.fi.uba.ar/6644/info/anespec/avanzado/real%20time/fundamentals%20of%20RTSA.pdf (Accessed: 12 August 2011).

Bradbury, J. (2000) '*Linear Predictive Coding'*, Yumpu,I Megazine, AG, Pp. 1-23, 5 December 2000. Available at: http://my.fit.edu/~vkepuska/ece5525/lpc_paper.pdf (Accessed: 11 July 2012).

Brookes, M., Huckwale, M., Naylor, P. (2012) 'Researching Speech Signal Enhancement. *Centre for Law Enforcement Audio Research'*, Clear Labs, Imperial College London, UK. Available at : http://www.ee.ic.ac.uk/hp/staff/dmb/courses/DSPDF/dspdf.htm (Accessed: 7 June 2011).

Byung-chul, J. (2001) '*Wavelet Transform Approach for Adaptive Filtering with Application to Fuzzy Neural Network Based Speech Recognition'*, PhD Dissertation, Wayne State University. Available at: http://elibrary.wayne.edu/record=b2823263~S47 (Accessed 17 january 2013).

Cai, Y., Ng, R. (2004) 'Indexing Spatio-Temporal Trajectories with Chebyshev Polynomials', *In Proceedings of ACM SIGMOD international conference on Management of data* (SIGMOD '04), Pp. 599-610, Paris, France, 13-18 June 2004.

Carlin, M. A., Thomas, S., Jansen, A., Hermansky, H. (2011) 'Rapid evaluation of speech representations for spoken term discovery', *in NTERSPEECH*, Pp. 821–824.

Carmi, A., Gurfil, P., Kanevsky , D. (2010) 'Methods for Sparse Signal Recovery Using Kalman Filtering With Embedded Pseudo-Measurement Norms and Quasi-Norms'. *IEEE Transactions on Signal Processing*, 58(4), Pp. 2405-2409.

Cassisi, C., Montalto, P., Aliotta, M., Cannata, A., Pulvirenti, A. (2012) '*Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining'*, Ln: Karahoca, A., Intech, Pp. 71-96.

Chan, C., Chan, C., Lee, L. (2014) 'Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity', *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Pp. 7814-7818, Florence, 4-9 May.

Chan, C., Lee, L. (2010) 'Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping', *Processing of Interspeech*, Pp. 693–696.

Chan, C., Lee, L. (2011) ‘Unsupervised hidden Markov modelling of spoken queries for spoken term detection without speech recognition’, *Processing of INTERSPEECH*, Pp. 2141–2144.

Chan, C., Lee, L. (2013) ‘Model-Based Unsupervised Spoken Term Detection with Spoken Queries’, *IEEE Trans on Audio, Speech, and Language Processing,* 21(7), Pp. 1330-1342.

Chan, K. P., Fu, A. W. C. (1999) ‘Efficient time Series Matching by Wavelets’, *Proc. of 15th International Conference on Data Engineering*, Pp. 126-133, Sydney, 23-26 March.

Chen, G. (2014) ‘*Low Resource Keyword Spotting’*, Department of Electrical and Computer Engineering, Johns Hopkins University. Available at: http://www.clsp.jhu.edu/~guoguo/papers/thesis_proposal.pdf (Accessed: 21 December 2013).

Chen, J., Benesty, J., Huang, Y., Doclo, S. (2006) ‘New Insights into the Noise Reduction Wiener Filter’, *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), Pp.1218-1234.

Dave, N. (2013) ‘Feature Extraction Methods LPC, PLP, and MFCC in Speech Recognition’, *International Journal for Advance Research in Eng. and Tech*, 01(06), Pp.1-5.

Dehak, N., Dehak, R., Glass, J., Reynolds, D., Kenny, P. (2010) ‘Cosine Similarity Scoring without Score Normalization Techniques’, *in Proc. Odyssey: The speaker and Language Recognition Workshop*, Brno, Czech Rebublic, 3-7 June .

Deng, L., O’Shaughnessy, D. (2003) ‘Speech Processing: A Dynamic and Optimization Oriented Approach’, *CRC Press, Signal processing and communication series*, 1st Edition, Pp. 29-63.

Dhingra, S., Nijhawan, G., Pandit, P. (2013) ‘Isolated Speech Recognition Using MFCC and DTW’, *Int. Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, 2(8), Pp. 1-8.

Djigan, V. I., Sovka, P., Cmejla, R. (1999) ‘Modified Spectral Subtraction Based Speech Enhancement’, *Proc. of the IEEE Workshop on Acoustics Echo and Noise Control*, IWAENC’99, Pp. 64-67.

Dudley, H. (1939) ‘The Vocoder’, *Bell Labs Record*, 17, Pp. 122-126.

Dudley, H., Riesz, R. R., Watkins, S. A. (1939) A Synthetic Speaker*, Journal of the Franklin Institute*, 227, Pp. 739-764.

Ephraim, Y. & Malah, D. (1984) ‘Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator’, *IEEE Transactions on Acoustics Speech and Signal Processing*, 32(6), Pp. 1109-1121.

Ephraim, Y., Malah, D. (1985) ‘Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator’, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2), Pp. 443-445.

Ephraim, Y., Malah, D., Juang, B. (1989) ‘On the Application of Hidden Markov Models for Enhancing Noisy Speech’, *IEEE Transactions on Acoustics, Speech and Signal Processing,* 37(12), Pp. 1846-1856.

Ephraim, Y., Van, T. H. (1995) ‘A Signal Subspace Approach for Speech Enhancement’, *IEEE Transactions on Speech and Audio Processing*, 3(4), Pp. 251-266.

Faloutsos, C., Ranganthan, M., Manolopoulos, Y. (1994) ‘Fast Subsequence Matching in Time Series Databases’, *ACM Proceedings of the SIGMOD International Conference on Management of Data*, New York, Pp. 419-429.

Farah, S., Shamim, A. (2013) ‘Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization’, *International Conference on Computer, Control & Communication* (IC4), Pp. 1-5, Karachi, 25-26 September.

Festvox CMU Arctic Databases (2007). ‘*Speech Synthesis Databases’*. Available at : http://festvox.org/index.html (Accessed: 12 July 2012).

Fing, Z., Lu, N., Jiang, P. (2008) ‘Posterior Probability Measure for Image Matching’, *International Journal of Pattern Recognition*, ELSEVIER. Vol. 41, Pp. 2422-2433.

Foley, B. G. (2012) ‘*A Dempster-Shafer Method for Multi-Sensor Fusion’*, MSc Thesis, Department of Mathematics and Statistics, Air Force Institute of Technology. Available at: http://www.dtic.mil/dtic/tr/fulltext/u2/a557749.pdf (Accessed: 02 December 2013).

Fousek, P., Hermansky, H. (2006) ‘Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition’, *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Vol. 1.

Fugal, D. L. (2009) ‘Conceptual Wavelets in Digital Signal Processing’. *Space and Signals Technologies LLC*, 47(2), Pp. 2-30.

Gabrea, M. (2001) ‘Adaptive Kalman Filtering-Based Speech Enhancement Algorithm’, *Electrical and Computer Engineering Conference* Vol. 1, Pp. 521- 526, Toronto, 13-16 May.

Galli, S., Logvinov, O. (2008) ‘Recent Developments in the Standardization of Power Line Communications within the IEEE’, *IEEE Communications Magazine,* 46(7), Pp. 64–71.

Gamulkiewicz, B., Weeks, M. (2003) ‘Wavelet based speech recognition’, *IEEE 46th Midwest Symposium on Circuits and Systems*, Vol.2, Pp. 678-681, Cairo, 30th December.

Gandhiraj, R., Sathidevi, P. S. (2007) ‘Auditory-based Wavelet Packet Filterbank for Speech Recognition using Neural Network’, *Proceedings of the 15th International Conference on Advanced Computing and Communications*, Pp. 666-671, Assam, 81-21 December.

Gannot, S., Burshtien, D., Weinstein, E. (1998) ‘Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms’, *IEEE Transactions; Speech and Audio Processing*, 6(4), Pp. 373-385.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V. (1993) ‘*TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1’*, Linguistic Data Consortium. Available at: https://catalog.ldc.upenn.edu/LDC93S1.

Gelb, A. (1974) ‘Applied Optimal Estimation’. Cambridge, MA: MIT Press, Pp. 72-79.

Giannakopoulos, T. (2014) ‘*A method for silence removal and segmentation of speech signals, implemented in Matlab*’. Available at: http://cgi.di.uoa.gr/~tyiannak/Software.html (Accessed: 13 May 2014).

Gibson, J. D., Koo, B., Gray, S. D. (1991) 'Filtering of Colored Noise for Speech Enhancement and Coding', *IEEE Transactions on Signal Processing*, 39(8), Pp. 1732-1742.

Gold, B., Rabiner, L. (1969) 'Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain', *Acoustical Society of America Journal*, 46(2), Pp. 442-448.

Grewal, M. S., Andrews, A. P. (1993) '*Kalman Filtering: Theory and Practice'*, *Prentice-Hall, Inc*., Upper Saddle River, NJ, USA.

Grewal, M. S., Andrews, A. P. (2001) '*Kalman Filtering Theory and Practice Using MATLAB*', 2nd Edition, Wiley & Sons, Pp. 15-17.

Griffin, D. W., Lim, J. S. (1984) 'Signal Estimation from Modified Short-Time Fourier Transform', *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2), Pp. 236-243.

Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M., Niblack, W. (1995) 'Efficient Color Histogram Indexing for Quadratic Form Distance Functions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7), Pp. 729-736.

Hagen, A., Pellom, B., Cole, R. (2007) 'Highly Accurate Children's Speech Recognition for Interactive Reading Tutors using Sub-word Units', *Speech Communications*, Elsevier, 49(12), Pp. 861-873.

Hand, D. J., Till, R. J. (2001) 'A Simple Generalization of the Area Under the ROC Curve to Multiple Class Classification Problems', *Machine Learning*, 45, Pp. 171-186.

Hazen, T. J., Shen, W., White, C. (2009) 'Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates', *IEEE Proceedings of the Automatic Speech Recognition & Understanding* (ASRU) Workshop, Pp. 421–426.

Hermansky, H. (1990) 'Perceptual Linear Predictive (PLP) Analysis of Speech', *Acoustical Society of America Journal*, 87(4), Pp.1738–1752.

Hermansky, H., Hanson, B., Wakita, H. (1985) 'Perceptually Based Linear Predictive Analysis of Speech', *IEEE Proceedings of Int. Conf. On Acoustic, Speech, and Signal processing*, Vol. 10, Pp. 509-512, April 1985.

Hirsch, H. G., Hellwing, K., Dobler, S. (2001) 'Speech Recognition at Multiple Sampling Rates', *Eurospeech,7th European Conference on Speech Communication and Technology*, Pp. 1-4, Aalborg, Denmark, 3-7 September.

Huijbregts, M., McLaren, M., Leeuwen, D. V. (2011) 'Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection', *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Pp. 4436–4439.

Hung, H., Chittaranjan, G. (2010) '*The Idiap Wolf Corpus: Exploring Group Behaviour in A Competitive Role-Playing Game*', ACM Multimedia, Italy. Available at: http://homepage.tudelft.nl/3e2t5/mmsct22567-hung.pdf (Accessed: 27 January 2011).

Jansen, A., Church, K., Hermansky, H. (2010) 'Towards spoken term discovery at scale with zero resources', Annual Conference of the International Speech Communication Association, Chiba, Japan, September 26-30.

Jansen, A., Durme, B. V. (2012) 'Indexing raw acoustic features for scalable zero resource search, in Proc. Of ICSA, INTERSPEECH, Pp. 2466-2469.

Jensen, J. H., Christensen, M. G., Daniel, P. W. E., Jensen, S. H. (2009) 'Quantitative Analysis of a Common Audio Similarity Measure', *IEEE Transactions on Audio, Speech, and Language Processing*, 17(04), Pp. 693-703.

Johnson, M. H., Alwan, A. (2003) 'Speech Coding: Fundamentals and Applications', *Wiley, Encyclopaedia of Telecommunications*, DOI: 10. 1002/04771219282.eot156.

Joho, H., Kishida, K. (2014) 'Overview of the NTCIR-11 Spoken Query & Doc task', *in Proceedings of the 11th NTCIR Conference*, Pp. 1–7, Tokyo, Japan, December 9-12.

Joukhadar, A., Scheuer, A., Laugier, C. (1999) 'Fast Contact Detection between Moving Deformable Polyhedral', *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, Kyongju, 17-21 October, Vol. 3, Pp. 1810–1815,

Junkawitsch, J., Neubauer, L., H¨oge, H., Ruske, G. (1996) 'A New Keyword Spotting Algorithm with Pre-calculated Optimal Thresholds', *Proceedings of the 4th*

International Conference on Spoken Language Processing (ICSLP-96), Philadelphia*.*

Juricka, M. (2014) '*Linear Predictive Speech Processing*' KTK, Department of Technical Cybernetics, Chapter 7, Pp. 1-25. Available at: http://frtk.fri.uniza.sk/jurecka/lpc2.pdf (Accessed: 02 February 2013).

Katsamanis, A., Papandreou, G., Maragos, P. (2009) 'Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation', *IEEE Transactions on Audio, Speech, And Language Processing*, 17(3), Pp. 411-422.

Kaur, A., Singh, T. (2010) 'Segmentation of Continuous Punjabi Speech Signal into Syllables', WCECS 2010, *Proceedings of the World Congress on Engineering and Computer Science*, Vol. 1, San Francisco, USA, 20-22 October.

Keogh, E. (2002) 'Exact Indexing of Dynamic Time Warping', *In Proceedings of the 28th International Conference on Very Large Data Bases,* Endowment, Hong Kong, Pp. 406-417.

Khan, W., Holton, R. (2015) 'Time Warped Continuous Speech Signal Matching Using Kalman Filter', *International Journal of Speech Technology*, Springer, 18(1), Pp. 1381-2416.

Khan, W., Jiang, P., Bilal, M. (2014) 'A Time Effective Approach for Vector Similarity Measurement Based on Vector Addition', *International Conference on Power, Control and VLSI Engineering* (ICPCVE'2014), Pp. 21-27, Dubai, 3-5 January.

Khan, W., Jiang, P., Chan, P. M. L. (2012) 'Word Recognition in Continuous Speech with Background Noise Based on Posterior Probability Measure', *IEEE international Conference on Electro/Information Technology*, Pp. 1-7, USA, 6-8 May.

Khan, W., Jiang, P., Chan, P. M. L. (2014) 'A Creative Application of Wavelet Transform and Kalman Filter for Children Proof-reading and Continuous Speech Tracking in Online Stories and TV Programs', *International Journal of Creative Computing*, Inderscience Publishers, 1(1), DOI: IJCRC-76367.

Khan, W., Jiang, P., Holton, R. (2014) 'Word Spotting in Continuous Speech Signal using Wavelet Transform'*, IEEE international Conference on Electro/Information Technology,* Pp. 275-279, Milwaukee, USA, 5-7 June.

Kuphaldt, T. R. (2007) ‘Lessons in Electric Circuits’, *Design Science License*, Vol. 4, Digital, 4th Edition, Pp. 407-427.

Lakshmikanth, S., Natraj, K. R., Rekha, K. R. (2014) ‘Noise Cancellation in Speech Signal Processing-A Review’, *Int. Jour. of Advanced Research in Computer and Communication Engineering*, 3(1), Pp. 5175-5186.

Li, J., Deng. L., Gong, Y., Haeb-Umbach, R. (2014) ‘An Overview of Noise-Robust Automatic Speech Recognition’, *Audio, Speech, and Language Processing, IEEE/ACM Transactions*, Vol. 22(4), Pp. 745-777.

Lieberman, P., Blumstien, S. E. (1988) ‘Speech Physiology, Speech Perception, and Acoustic Phonetics’, *Cambridge University Press, Language Arts & Disciplines*, 4th February, Pp. 3-39.

Lin, J. (1991) ‘Divergence Measures Based on the Shannon Entropy’, *IEEE Transactions on Information Theory*, 37(1), Pp. 145–151.

Lin, Y. S., Ji, C. P. (2010) ‘Research on Improved Algorithm of DTW in Speech Recognition’, *International Conference on Computer Application and System Modelling*, Taiyuan, 22-24 October, Vol. 9, Pp. 418-421.

Lipeika, A. (2010) ‘Optimization of Formant Feature Based Speech Recognition’, *Institute of Mathematics and Informatics, Informatica*, 21(3), Pp. 361-374.

Liscombe, M., Asif, A. (2009) ‘A new method for instantaneous signal period identification by repetitive pattern matching’, *IEEE, International Multi-topic Conference* (INMIC), Pp. 1-5, Islamabad, 14-15 December.

Liu, T. L., Chen, H. T. (2004) ‘Real-Time Tracking using Trust-Region Methods’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (3), Pp. 397–402.

Lyons, J. (2012) ‘*Mel Frequency Cepstral Coefficient (MFCC) Tutorial*’, Practical Cryptography. Available at: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ (Accessed 12 December 2012).

Maher, A.G., King, R.W.,Rathmell, J.G. (1992) ‘A Comparison of Noise Reduction Techniques for Speech Recognition in Telecommunications Environments’, *The*

Institution of Engineers Australia Communications Conference, Sydney, October 20-22 1992, Pp. 107-111.

Mandal, A., Kumar, K. R., Mitra, P. (2014) ‘Recent developments in spoken term detection: a survey’, *International Journal of Speech Technology*, 17(02), Pp. 183-198.

Martin, R., Cox, R. V. (1999) ‘New Speech Enhancement Techniques for Low Bit Rate Speech Coding’, *Proceedings of IEEE Workshop on Speech Coding*, Pp. 165-167.

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matějka, P., Černocký, J., Poh, N., Kittler, J., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J., Tresadern, P., Cootes, T. (2012) ‘Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data’, *IEEE International Conference on Multimedia and Expo Workshops,* Pp. 635-640.

McGee, L. A., Schmidt, S. F. (1985) ‘Discover of the Kalman Filter as a Practical Tool for Aerospace and Industry’, *NASA Technical Memorandum,* 86847, Pp. 1-21.

Meinard, M. (2007) ‘*Information Retrieval for Music and Motion*’, Springer Berlin Heidelberg, Chapter 2, Pp 51-67.

Metze, F., Barnard, E., Davel, M., Heerden, C., Anguera, X., Gravier, G., Rajput, N. (2012) ‘Spoken web search CEUR Workshop Proceedings’, *in Proceedings of MediaEval,* Aachen, Germany, Pp. 1–2.

Meyer, Y. (1989) ‘*Wavelets and Applications*’, Masson & Springer, Verlag, Pp. 61-68.

Michael, E. (2001) ‘*Top 100 speeches*’, American Rhetoric. Available at: http://www.americanrhetoric.com/top100speechesall.html (Accessed: 12 December 2013).

MIT (2003) ‘*Dynamic Time Warping & Search*’, Lecture 9. Available at: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture9.pdf (Accessed: 13 March 2011).

Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B. (2009) '*Exact Discovery of Time Series Motifs', University of California*. Available at: http://www.cs.ucr.edu/~eamonn/exact_motif/EM.pdf (Accessed: 13 January 2012).

Muscariello, A., Gravier, G., Bimbot, F. (2009) 'Audio keyword extraction by unsupervised word discovery*', Annual Conference of the International Speech Communication Association,* INTERSPEECH, Brighton, UK, Pp. 2843-2846, 1-5 September.

Myers, C.S., Rabiner, L.R. (1981) 'A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition', *The Bell System Technical Journal*, 60(7), Pp. 1389-1409.

Nortel, N. (2002) '*Voice Fundamentals: From Analogue to ATM'*, Nortel Networks Corporation, Pp. 1-4.

Online Audio Stories. (2008) '*Short Stories and Audio Books Online for Kids*', Available at: http://www.onlineaudiostories.com/category/all_stories/audio_stories/ (Accessed 01Feb 2013).

Parada, C., Sethy, A., Ramabhadran, B. (2009) 'Query-By-Example Spoken Term Detection for OOV Terms', *IEEE Proceedings of the Automatic Speech Recognition & Understanding* (ASRU) Workshop. Pp. 404–409.

Park, A., Glass, J. R. (2008) 'Unsupervised pattern discovery in speech', *IEEE Transactions on Audio, Speech, and Language Processing*, 16(01), Pp. 186–197.

Peinado, A. M., Lopez, J. M., Sanchez, V. E., Segura, J. C., Ayuso, A. J. R. (1991) 'Improvements in HMM based Isolated Word Recognition System', *IEEE Proceedings in Communications*, Speech and Vision, 138(3), Pp. 201-206.

Policar, R. (1996) '*The Engineer's Ultimate Guide to Wavelet Analysis: Wavelet Tutorial*', Available at: http://person.hst.aau.dk/enk/ST8/wavelet_tutotial.pdf (Accessed: 23 October 2012).

Policar, R. (2006) '*The Wavelet Tutorial'*, Rowan University, College of Engineering. Available at: http://users.rowan.edu/~polikar/WAVELETS/WTpart4.html (Accessed: 12 M ay 2012).

Portnoff, M. R. (1981) ‘Time-Scale Modification of Speech Based on Short-Time Fourier Analysis’, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3), Pp. 374-390.

Pour, M. M., Farokhi, F. (2009) ‘An Advanced Method for Speech Recognition’, *World Academy of Science, Engineering and Technology*, 3(1), Pp. 840-845.

Rabiner, L. R., Juang, B. H. (1993) ‘Fundamental of Speech Recognition’, *PTR Prentice-Hall*, New Jersey, Pp. 37-51.

Rabiner, L. R., Juang, B. H. (2004) ‘*Statistical Methods for the Recognition and Understanding of Speech*’, Encyclopedia of Language and Linguistics. Available at: http://www.ece.ucsb.edu/faculty/Rabiner/ece259/Reprints/355_Statistical%20Methods%20for%20ASR-final-1.pdf (Accessed: 09 April 2012).

Rabiner, L. R., Sambur, M.R. (2013) ‘An Algorithm for Determining the Endpoints of Isolated Utterances’, *The Bell System Technical Journal*, 54(2), Pp. 297-315.

Rabiner, L. R., Schafer, R.W. (1978) ‘Digital Processing of Speech Signals’, *Prentice-Hall signal processing series*, Englewood Cliffs, Prentice-Hall. Pp. 117-221.

Ratanamahatana, C. A., Keogh, E. (2005) ‘Three Myths about Dynamic Time Warping Data Mining’, *Proceeding of SIAM International Conference on Data Mining*, Pp. 506-510.

Ravindran,G., Shenbagadevi, S., Selvam, V. S. (2010) ‘Cepstral and Linear Prediction Techniques for Improving Intelligibility and Audibility of Impaired Speech’, *Journal of Biomedical Science and Engineering*, 3(1), Pp. 85-94.

Renger, B., Feng, J., Dan, O., Chang, H., Barbosa, L. (2011) ‘Voice-Enabled Social TV’, *Proceedings of the 20th international conference companion on World wide web,* Pp. 253-256, India, 28 March- 1 April.

Rodman, J. (2006) ‘*The Effect of Bandwidth on Speech Intelligibility’*. Policom, Pp. 2-9. Available at: http://www.ivci.com/pdf/whitepaper-polycom-bandwith-and-speech-intelligibility.pdf (Accessed: 29 October 2011).

Rui, Y., Chen, Y. (2001) ‘Better Proposal Distributions: Object Tracking using Unscented Particle Filter’, *In Processing, IEEE Conference on Computer Vision and Pattern Recognition,* Kauai, Hawaii, 8-14 December, Vol 2, Pp. 786–793.

Ruzanski, E., Hansen, J. H. L., Meyerhoff, J., Saviolakis, G., Norris, W., Wollert, T. (2006) ‘Stress Level Classification of Speech Using Euclidean Distance Metrics in a Novel Hybrid Multi-Dimensional Feature Space’, ICASSP Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, 14-16 May, Vol.1, Pp. 1-1, doi: 10.1109/ICASSP.2006.1660048.

Saha, G., Chakroborty, S., Senapati, S. (2005) ‘A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications’, *Proceedings of the NCC,* 2-5 January.

Sahoo, T. R., Patra, S. (2014) ‘Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification’, *Int. Jour. Image, Graphics and Signal Processing*, (06), Pp. 27-35.

Sakoe, H. (2003) ‘Two-level DP-Matching, a Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition’, *IEEE Transaction on Acoustics, Speech and Signal Processing*, 627(6), Pp. 588-595.

Sakoe, H., Chiba, S. (1978) ‘Dynamic programming algorithm optimization for spoken word recognition’, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1), Pp. 43– 49.

Sarikaya, R. (2001) ‘*Robust and Efficient Techniques for Speech Recognition in Noise*’, PhD Dissertation, Duke University. Available at: http://dl.acm.org/citation.cfm?id=934525 (Accessed: 22 August 2012).

Sarma, V. V. S., Venugopal, D. (1978) ‘Studies on Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification’, *IEEE International Conference on ICASSP*, Vol. 3, Pp. 1-4, April 1978.

Shafik, A., Elhalafawy, S. M., Diab, S.M., Sallam, B. M., Abd El-samie, F. E. (2009) ‘A Wavelet Based Approach for Speaker Identification from Degraded Speech’, *International Journal of Communication Networks and Information Security*, 1(3), Pp. 52-58.

Shamoo, A. E., Resnik, D. B. (2003) ‘*Responsible Conduct of Research’*, New York. Oxford University Press. Available at: http://ir.nmu.org.ua/bitstream/handle/123456789/126128/06c4f6a7e41bce1851b87a758068b680.pdf?sequence=1 (Accessed: 9th December 2015).

Sharma, P., Rajpoot, A. K. (2013) 'Automatic Identification of Silence, Unvoiced and Voiced Chunks in Speech', *Jounal of Computer Science & Information Technology*, 03(05), Pp. 87-96.

Sher, M., Ahmad, N., Sher, M. (2012) 'TESPAR Feature Based Isolated Word Speaker Recognition System', *Proceedings of the 18th International Conference on Automation & Computing*, United Kingdom, 7-8 September, Pp. 1-4.

Shuttleworth, M. (2008) '*Quantitative Research Design'*, Exploreable.com. Available at: https://explorable.com/quantitative-research-design (Accessed: 15th January 2015).

Soluade, O. A. (2010) 'Establishment of Confidence Threshold for Interactive Voice Response Systems Using ROC Analysis', *Communications of the IIMA*, 10(2), Pp. 43-57.

Tejedor, J., Toledano, D. T., Anguera, X., Varona, A., Hurtado, L. F., Miguel, A., Colas, J. (2013) 'Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion', *Journal of Audio, Speech, and Music Processing,* EURASIP, (23), Pp. 1–17.

Tejedor, J., Toledano, D. T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal, A., Echeverry-Correa, J. D., Coucheiro-Limeres, A., Olcoz, J., Miguel, A. (2015) 'Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion', *EURASIP Journal on Audio, Speech, and Music Processing,* (21), Pp. 1-27.

Thakur, A. S., Sahayam, N. (2013) 'Speech Recognition Using Euclidean Distance', *International Journal of Emerging Technology and Advanced Engineering*, 3(3), Pp. 587-590.

Thambiratmann, K., Sridharan, S. (2007) 'Rapid yet accurate speech indexing using dynamic match lattice spotting', *IEEE Trans. Audio Speech Lang. Process*. 15(1), Pp. 346–357.

Theodoridis, S., Koutroumbas, K. (2003) '*Pattern Recognition', Academic Press,* New York, Pp. 337–340.

Tyagi, V., Wellekens, C. (2005) 'On Desensitizing the Mel-Cepstrum to Spurious Spectral Components for Robust Speech Recognition', *IEEE International*

Conference on, Acoustics, Speech, and Signal Processing, 18-23 March, Vol. 1, Pp. 529–532.

Wang, H., Lee, T., Leung, C. (2011) Unsupervised Spoken Term Detection with Acoustic Segment Model', *IEEE Proceedings of the International Conference on Speech Database and Assessments* (Oriental COCOSDA), Pp. 106–111.

Wang, Y. (2006) '*Voice and Audio Digitization and Sampling Rate Conversion*, Lab Manual, Polytechnic University, Brooklyn. Available at: http://eeweb.poly.edu/~yao/EE3414/sampling.pdf (Accessed: 22 December 2011).

Willet, P., Barnard, J. M., Downs, G. M. (1998) 'Chemical Similarity Searching', *Journal of Chemical Information and Computer Science*, 38(6), Pp. 983-996.

Yegnanarayana, B., Sreekumar, T. (1984) 'Signal Dependent Matching for Isolated Word Speech Recognition System', *Journal of Signal Processing*, 7(2), Pp. 161-173.

Young, S. (2008) '*HMMs and Related Speech Recognition Technologies*', Springer Handbook of Speech Processing, Springer-Verlag, Heidelberg, Berlin, Pp. 539-583.

Yuan, S. T., Sun, J. (2005) 'Ontology-based structured cosine similarity in document summarization: with applications to mobile audio-based knowledge management', *IEEE Trans. Syst. Man. Cybern. B. Cybern*, Pp. 1028-40.

Zahorian, S. A., Hu, H. (2008) 'A Spectral/Temporal Method for Robust Fundamental Frequency Tracking', *Journal of Acoustic Society of America*, 123(6), Pp. 4559-4571.

Zhang, S. (2006) 'An energy-based adaptive voice detection approach', *International Conference on Signal Processing*, Beijing, Vol. 01, 16-20 November.

Zhang, Y., Adl, K., Glass, J. R. (2012) 'Fast spoken query detection using lower-bound dynamic time warping on graphical processing units', *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Pp. 5173–5176, Kyoto, 25-30 March.

Zhang, Y., Glass, J. R. (2009) 'Unsupervised Spoken Keyword Spotting Via Segmental DTW on Gaussian Posteriorgrams', *IEEE Proceedings of the Automatic Speech Recognition & Understanding* (ASRU) Workshop, Pp. 398–403.

Zhang, Y., Glass, J. R. (2010) 'Towards multi-speaker unsupervised speech pattern discovery', *IEEE, Int. Conf. on Acoustics Speech and Signal Processing*, Pp. 4366-4369, Dallas, TX, 14-19 March.

Zhang, Y., Glass, J. R. (2011) 'A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping', *Processing of ISCA,* INTERSPEECH, Pp. 1909–1912, Florence, Italy, 28-31 August.

Zhang, Y., Glass, J. R. (2011) 'An Inner-Product Lower-Bound Estimate for Dynamic Time Warping', *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Pp. 5660–5663.

[2003] '*Dynamic Time Warping & Search*', MIT Lectures No. 9. Available at: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture9.pdf (Accessed: 21 December 2011).

[no date] '*Speech Enhancement'*. Available at: http://www.cs.tut.fi/sgn/arg/8003051/ehostus_en.pdf (Accessed: 12 April 2012).

APPENDICES

Appendix A

➢ Peer Reviewed Publications

I. Khan, W., Holton, R. (2015) ‘Time Warped Continuous Speech Signal Matching Using Kalman Filter’, *International Journal of Speech Technology*, Springer, 18(1), Pp. 1381-2416.

II. Khan, W., Jiang, P., Chan, P. M. L. (2014) ‘A Creative Application of Wavelet Transform and Kalman Filter for Children Proof-reading and Continuous Speech Tracking in Online Stories and TV Programs’, *International Journal of Creative Computing*, Inderscience Publishers, 1(1), DOI: IJCRC-76367.

III. Khan, W., Holton, R. (2015) ‘Adaptive Framing Based Similarity Measurement between Time Warped Speech Signals Using Kalman Filter’, *EURASIP Audio Journal on Audio, Speech, and Music Processing*, Springer (Under Review).

IV. Khan, W., Holton, R. (2016) ‘Decision Support System for Keyword Spotting Using Theory of Evidence’, *International Conference on Information Management and Industrial Engineering* (ICII), Los Angeles, 4-6 January.

V. Khan, W., Jiang, P., Holton, R. (2014) ‘Word Spotting in Continuous Speech Signal using Wavelet Transform’*, IEEE international Conference on Electro/Information Technology,* Pp. 275-279, Milwaukee, USA, 5-7 June.

VI. Khan, W., Jiang, P., Bilal, M. (2014) ‘A Time Effective Approach for Vector Similarity Measurement Based on Vector Addition’, *International Conference on Power, Control and VLSI Engineering* (ICPCVE'2014), Pp. 21-27, Dubai, 3-5 January.

VII. Khan, W., Jiang, P., Chan, P. M. L. (2012) ‘Word Recognition in Continuous Speech with Background Noise Based on Posterior Probability Measure’, *IEEE international Conference on Electro/Information Technology*, Pp. 1-7, USA, 6-8 May.

➢ Publications as Co-Author

I. Bilal, M., Chan, P. M. L., Khan, W. (2013) ‘Cooperative Network for Emergency Communications: Fair Distribution of Reward among Players Based on Their Marginal Contribution’, *Cyber Journals: Multidisciplinary Journals in Science and Technology (JSAT),* 3(8), Pp. 11-25.

II. Bilal, M., Khan, W., Yar, A., Mockford, S., Awan, I. U. A. (2012) **‘**Tracesaver: A Tool for Network Service Improvement and Personalized Analysis of User Centric Statistics’, *PCGlobal conference*, Pp. 215-221, Las Vegas,USA, 6-8 August.

III. Badii, A., Khan, A. A., Khan, W. (2014) ‘Situation Assessment through Multi-modal Sensing of Dynamic Environments to Support Cognitive Robot Control’, *FACTA UNIVERSITATIS* Series: Mechanical Engineering, 12(2), Pp. 251-260.

Appendix B: Detailed Results

Table 1: Statistical Results for Adaptive Speech Tracking Based on Different Approaches

Tracking Approaches	**KF Based Adaptive Tracking**	**Similarity Measurement Approaches**			
		Evaluation Metrics	**Mean-MFCC**	**DTW**	**DTWC**
		Sensitivity	0.9918	0.8999	0.7334
		Specificity	0.9726	0.8702	0.9949
		Matching Accuracy	0.9801	0.8958	0.9455
		1/LR+	0.0276	0.1601	0.2457

		LR-		0.0084	0.6718	0.9612
		F-Score		0.9713	0.4357	0.0942
		Tracking Accuracy (%)		1	0.9484	0.8914
		Avg. Execution Time (seconds)		1.3747	3.0419	1.6250
		Type I Error	μ	0.0011	0.0016	3.0613 e-06
			σ	0.0016	0.0079	1.3315 e-05
		Type II Error	μ	3.1270e-04	0.5258	0.9254
			σ	8.8200e-04	0.3391	0.0862
	SR Based Non- Adaptive Tracking	Sensitivity		0.7299	0.7517	0.1399
		Specificity		0.9856	0.9421	0.9998
		Matching Accuracy		0.9415	0.9121	0.8739
		LR+		0.0182	0.0909	0.0046
		LR-		0.2717	0.2560	0.8603
		F-Score		0.8957	0.8698	0.2587
		Tracking Accuracy (%)		0.6822	0.6991	0.1515
		Type I Error	μ	0.0011	0.0047	1.1904 e-06
			σ	0.0016	0.0052	5.2907 e-06
		Type II Error	μ	3.1270e-04	0.1462	0.7569
			σ	8.8200e-04	0.2237	0.2035

Table 2: Statistical Results for Speech Tracking Based on Euclidean Distance and Mean MFCC

Kalman Filter Based Adaptive Speech Tracking & Similarity Measurement (ED+Mean_MFCC)											
Test Cases	TP	TN	FP	FN	Sen	Spec	Acc	LR+	LR-	F-Score	Tracking Accuracy
1	25	57	0	0	1	1	1	0	0	1	1
2	17	39	0	0	1	1	1	0	0	1	1
3	43	92	3	0	1	0.97	0.98	0.03	0	0.97	1
4	32	71	3	0	1	0.96	0.97	0.04	0	0.96	1
5	45	59	4	0	1	0.94	0.96	0.06	0	0.96	1
6	27	66	1	0	1	0.99	0.99	0.01	0	0.98	1
7	26	54	1	0	1	0.98	0.99	0.02	0	0.98	1
8	30	64	2	0	1	0.97	0.98	0.03	0	0.97	1
9	26	78	0	1	0.96	1	0.99	0	0.04	0.98	1
10	157	140	14	0	1	0.91	0.95	0.09	0	0.96	1
11	88	139	6	0	1	0.96	0.97	0.04	0	0.97	1
12	77	140	1	0	1	0.99	1	0.01	0	0.99	1
13	65	87	1	0	1	0.99	0.99	0.01	0	0.99	1
14	59	111	2	2	0.97	0.98	0.98	0.02	0.03	0.97	1
15	93	180	2	0	1	0.99	0.99	0.01	0	0.99	1
16	71	91	9	1	0.99	0.91	0.94	0.09	0.02	0.93	1
17	62	131	3	0	1	0.98	0.98	0.02	0	0.98	1
18	87	160	5	0	1	0.97	0.98	0.03	0	0.97	1
19	78	118	3	0	1	0.98	0.98	0.02	0	0.98	1
20	43	87	3	0	1	0.97	0.98	0.03	0	0.97	1
21	86	215	6	1	0.99	0.97	0.98	0.03	0.01	0.96	1
22	200	409	8	0	1	0.98	0.99	0.02	0	0.98	1
23	129	227	13	1	0.99	0.95	0.96	0.05	0.01	0.95	1
24	119	229	5	1	0.99	0.98	0.98	0.02	0.01	0.98	1
25	108	245	2	0	1	0.99	0.99	0.01	0	0.99	1
26	104	154	4	0	1	0.97	0.98	0.03	0	0.98	1
27	109	265	6	1	0.99	0.98	0.98	0.02	0.01	0.97	1
28	143	309	8	3	0.98	0.97	0.98	0.03	0.02	0.96	1

29	104	262	4	2	0.98	0.98	0.98	0.02	0.02	0.97	1
30	132	259	5	0	1	0.98	0.99	0.02	0	0.98	1
31	195	422	7	0	1	0.98	0.99	0.02	0	0.98	1
32	91	225	2	6	0.94	0.99	0.98	0.01	0.06	0.96	1
33	65	160	2	0	1	0.99	0.99	0.01	0	0.98	1
34	108	189	4	4	0.96	0.98	0.97	0.02	0.04	0.96	1
35	118	265	7	7	0.94	0.97	0.96	0.03	0.06	0.94	1
36	86	155	7	0	1	0.96	0.97	0.04	0	0.96	1
37	93	190	2	0	1	0.99	0.99	0.01	0	0.99	1
38	32	68	0	0	1	1	1	0	0	1	1
39	127	147	7	1	0.99	0.95	0.97	0.05	0.01	0.97	1
40	88	136	4	0	1	0.97	0.98	0.03	0	0.98	1
41	109	184	7	0	1	0.96	0.98	0.04	0	0.97	1
42	64	104	4	0	1	0.96	0.98	0.04	0	0.97	1
43	86	191	7	1	0.99	0.96	0.97	0.04	0.01	0.96	1
44	86	188	5	0	1	0.97	0.98	0.03	0	0.97	1
45	28	71	4	2	0.93	0.95	0.94	0.06	0.07	0.90	1
46	121	225	5	2	0.98	0.98	0.98	0.02	0.02	0.97	1
47	84	178	6	2	0.98	0.97	0.97	0.03	0.02	0.95	1
48	132	224	5	0	1	0.98	0.99	0.02	0	0.98	1
49	48	127	3	0	1	0.98	0.98	0.02	0	0.97	1
50	98	192	4	0	1	0.98	0.99	0.02	0	0.98	1
51	99	176	6	1	0.99	0.97	0.98	0.03	0.01	0.97	1
52	29	87	2	0	1	0.98	0.98	0.02	0	0.97	1
53	127	190	7	0	1	0.96	0.98	0.04	0	0.97	1
54	101	150	4	0	1	0.97	0.98	0.03	0	0.98	1
55	127	234	5	0	1	0.98	0.99	0.02	0	0.98	1
56	87	102	6	0	1	0.94	0.97	0.06	0	0.97	1
57	119	204	9	0	1	0.96	0.97	0.04	0	0.96	1
58	80	171	3	2	0.98	0.98	0.98	0.02	0.02	0.97	1
59	78	118	3	0	1	0.98	0.98	0.02	0	0.98	1
60	43	87	3	0	1	0.97	0.98	0.03	0	0.97	1

Table 3: Statistical Results for Non-Adaptive Speech Tracking Based on Euclidean Distance and Mean MFCC

Search Region Based Non-Adaptive Tracking & Similarity Measurement Performance (ED+Mean_MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	24	65	2	0	1	0.97	0.98	0.03	0	0.96	1
2	19	42	2	0	1	0.95	0.97	0.05	0	0.95	1
3	20	120	2	12	0.63	0.98	0.91	0.03	0.38	0.74	0.50
4	33	82	4	0	1	0.95	0.97	0.05	0	0.94	1
5	34	84	1	0	1	0.99	0.99	0.01	0	0.99	1
6	30	73	2	0	1	0.97	0.98	0.03	0	0.97	1
7	24	59	1	0	1	0.98	0.99	0.02	0	0.98	1
8	14	87	1	10	0.58	0.99	0.90	0.02	0.42	0.72	0.44
9	14	94	1	10	0.58	0.99	0.91	0.02	0.42	0.72	0.47
10	102	252	2	1	0.99	0.99	0.99	0.01	0.01	0.99	1
11	79	181	6	0	1	0.97	0.98	0.03	0	0.96	1
12	73	169	3	0	1	0.98	0.99	0.02	0	0.98	1
13	51	123	1	0	1	0.99	0.99	0.01	0	0.99	1
14	50	136	2	1	0.98	0.99	0.98	0.01	0.02	0.97	1
15	92	220	3	0	1	0.99	0.99	0.01	0	0.98	1
16	29	153	3	11	0.72	0.98	0.93	0.03	0.28	0.81	0.71
17	60	158	6	0	1	0.96	0.97	0.04	0	0.95	1
18	72	209	4	2	0.97	0.98	0.98	0.02	0.03	0.96	1
19	74	143	7	0	1	0.95	0.97	0.05	0	0.95	1
20	42	108	4	0	1	0.96	0.97	0.04	0	0.95	1
21	99	244	7	0	1	0.97	0.98	0.03	0	0.97	1
22	115	546	6	47	0.71	0.99	0.93	0.02	0.29	0.81	0.55
23	19	357	7	51	0.27	0.98	0.87	0.07	0.74	0.40	0.19
24	49	325	1	31	0.61	1	0.92	0.01	0.39	0.75	0.48
25	31	337	4	41	0.43	0.99	0.89	0.03	0.58	0.58	0.32
26	53	238	1	16	0.77	1	0.94	0.01	0.23	0.86	0.66
27	89	321	7	17	0.84	0.98	0.94	0.03	0.16	0.88	0.74

28	4	460	0	75	0.05	1	0.86	0	0.95	0.10	0.04
29	78	328	6	15	0.84	0.98	0.95	0.02	0.16	0.88	0.77
30	69	357	7	29	0.70	0.98	0.92	0.03	0.30	0.79	0.58
31	30	599	7	85	0.26	0.99	0.87	0.04	0.75	0.39	0.18
32	30	308	1	39	0.43	1	0.89	0.01	0.57	0.60	0.30
33	15	215	0	29	0.34	1	0.89	0	0.66	0.51	0.24
34	6	297	0	47	0.11	1	0.87	0	0.89	0.20	0.08
35	21	393	0	55	0.28	1	0.88	0	0.72	0.43	0.21
36	22	238	4	30	0.42	0.98	0.88	0.04	0.59	0.56	0.31
37	5	279	0	45	0.10	1	0.86	0	0.90	0.18	0.06
38	29	82	1	0	1	0.99	0.99	0.01	0	0.98	1
39	33	264	2	30	0.52	0.99	0.90	0.01	0.48	0.67	0.38
40	1	228	0	37	0.03	1	0.86	0	0.97	0.05	0.05
41	92	249	2	0	1	0.99	0.99	0.01	0	0.99	1
42	61	135	0	0	1	1	1	0	0	1.00	1
43	31	272	1	32	0.49	1	0.90	0.01	0.51	0.65	0.35
44	37	266	0	26	0.59	1	0.92	0	0.41	0.74	0.47
45	13	101	0	12	0.52	1	0.90	0	0.48	0.68	0.39
46	23	333	4	46	0.33	0.99	0.88	0.04	0.67	0.48	0.22
47	37	250	1	27	0.58	1	0.91	0.01	0.42	0.73	0.44
48	119	290	4	0	1	0.99	0.99	0.01	0	0.98	1
49	50	145	1	0	1	0.99	0.99	0.01	0	0.99	1
50	84	242	3	7	0.92	0.99	0.97	0.01	0.08	0.94	0.90
51	29	268	0	32	0.48	1	0.90	0	0.52	0.64	0.34
52	38	85	3	0	1	0.97	0.98	0.03	0	0.96	1
53	99	274	2	3	0.97	0.99	0.99	0.01	0.03	0.98	0.98
54	86	198	3	0	1	0.99	0.99	0.01	0	0.98	1
55	115	301	4	0	1	0.99	0.99	0.01	0	0.98	1
56	64	157	3	0	1	0.98	0.99	0.02	0	0.98	1
57	101	261	9	0	1	0.97	0.98	0.03	0	0.96	1
58	13	246	0	35	0.27	1	0.88	0	0.73	0.43	0.19
59	74	143	7	0	1	0.95	0.97	0.05	0	0.95	1

60	42	108	4	0	1	0.96	0.97	0.04	0	0.95	1

Table 4: Statistical Results for Speech Tracking with Traditional Silence Removal Approach

Kalman Filter with Energy & Spectral Centroid Silence Removal Approach (ED+Mean_MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	23	58	1	0	1	0.98	0.99	0.02	0	0.98	1
2	24	51	0	0	1	1	1	0	0	1	1
3	69	124	11	0	1	0.92	0.95	0.08	0	0.93	1
4	75	85	12	0	1	0.88	0.93	0.12	0	0.93	1
5	57	109	6	1	0.98	0.95	0.96	0.05	0.02	0.94	1
6	44	99	3	1	0.98	0.97	0.97	0.03	0.02	0.96	1
7	48	82	4	1	0.98	0.95	0.96	0.05	0.02	0.95	1
8	41	92	1	1	0.98	0.99	0.99	0.01	0.02	0.98	1
9	41	77	5	0	1	0.94	0.96	0.06	0	0.94	1
10	139	255	4	0	1	0.98	0.99	0.02	0	0.99	1
11	206	217	27	1	1	0.89	0.94	0.11	0.01	0.94	1
12	112	229	3	2	0.98	0.99	0.99	0.01	0.02	0.98	1
13	70	141	8	1	0.99	0.95	0.96	0.05	0.01	0.94	1
14	0	180	28	20	0	0.87	0.79	0.81	0.99	NaN	0.1
15	138	234	8	0	1	0.97	0.98	0.03	0	0.97	1
16	99	84	12	0	1	0.88	0.94	0.13	0	0.94	1
17	129	199	5	0	1	0.98	0.98	0.02	0	0.98	1
18	134	255	10	1	0.99	0.96	0.97	0.04	0.01	0.96	1
19	113	203	9	3	0.97	0.96	0.96	0.04	0.03	0.95	1
20	98	189	5	0	1	0.97	0.98	0.03	0	0.98	1
21	169	319	19	8	0.95	0.94	0.95	0.06	0.05	0.93	1
22	342	565	17	4	0.99	0.97	0.98	0.03	0.01	0.97	1
23	239	391	19	0	1	0.95	0.97	0.05	0	0.96	1
24	451	116	112	4	0.99	0.51	0.83	0.50	0.02	0.89	0.97
25	381	50	175	4	0.99	0.22	0.71	0.79	0.05	0.81	0.97
26	154	289	9	4	0.97	0.97	0.97	0.03	0.03	0.96	1

27	217	358	18	2	0.99	0.95	0.97	0.05	0.01	0.96	1
28	223	395	11	0	1	0.97	0.98	0.03	0	0.98	1
29	294	380	20	3	0.99	0.95	0.97	0.05	0.01	0.96	1
30	209	416	7	6	0.97	0.98	0.98	0.02	0.03	0.97	1
31	401	718	21	5	0.99	0.97	0.98	0.03	0.01	0.97	1
32	185	315	3	6	0.97	0.99	0.98	0.01	0.03	0.98	1
33	139	198	11	0	1	0.95	0.97	0.05	0	0.96	1
34	134	267	32	6	0.96	0.89	0.91	0.11	0.05	0.88	1
35	262	334	25	8	0.97	0.93	0.95	0.07	0.03	0.94	1
36	167	223	17	0	1	0.93	0.96	0.07	0	0.95	1
37	188	280	5	1	0.99	0.98	0.99	0.02	0.01	0.98	1
38	70	108	5	1	0.99	0.96	0.97	0.04	0.01	0.96	1
39	133	253	12	0	1	0.95	0.97	0.05	0	0.96	1
40	182	244	13	0	1	0.95	0.97	0.05	0	0.97	1
41	209	382	18	2	0.99	0.95	0.97	0.05	0.01	0.95	1
42	180	166	33	5	0.97	0.83	0.90	0.17	0.03	0.90	1
43	184	252	16	1	0.99	0.94	0.96	0.06	0.01	0.96	1
44	155	245	8	0	1	0.97	0.98	0.03	0	0.97	1
45	48	71	3	0	1	0.96	0.98	0.04	0	0.97	1
46	60	127	2	0	1	0.98	0.99	0.02	0	0.98	1
47	190	326	11	2	0.99	0.97	0.98	0.03	0.01	0.97	1
48	143	321	9	8	0.95	0.97	0.96	0.03	0.05	0.94	1
49	87	233	3	1	0.99	0.99	0.99	0.01	0.01	0.98	1
50	240	178	32	0	1	0.85	0.93	0.15	0	0.94	1
51	170	250	11	1	0.99	0.96	0.97	0.04	0.01	0.97	1
52	52	140	4	1	0.98	0.97	0.97	0.03	0.02	0.95	1
53	199	268	12	0	1	0.96	0.97	0.04	0	0.97	1
54	119	255	6	0	1	0.98	0.98	0.02	0	0.98	1
55	196	182	25	2	0.99	0.88	0.93	0.12	0.01	0.94	1
56	112	204	7	0	1	0.97	0.98	0.03	0	0.97	1
57	333	142	85	0	1	0.63	0.85	0.37	0	0.89	1
58	139	207	7	0	1	0.97	0.98	0.03	0	0.98	1

59	113	203	9	3	0.97	0.96	0.96	0.04	0.03	0.95	1
60	98	189	5	0	1	0.97	0.98	0.03	0	0.98	1

Table 5: Statistical Results for Speech Tracking Based on DTW and MFCC

Kalman Filter Based Adaptive Speech Tracking & Similarity Measurement (DTW+MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	13.00	55.00	1.00	5.00	0.72	0.98	0.92	0.02	0.28	0.81	1.00
2	2.00	50.00	0.00	7.00	0.22	1.00	0.88	0.00	0.78	0.36	0.50
3	42.00	80.00	3.00	3.00	0.93	0.96	0.95	0.04	0.07	0.93	1.00
4	10.00	83.00	0.00	9.00	0.53	1.00	0.91	0.00	0.47	0.69	1.00
5	14.00	71.00	3.00	10.00	0.58	0.96	0.87	0.07	0.43	0.68	0.50
6	22.00	66.00	2.00	4.00	0.85	0.97	0.94	0.03	0.16	0.88	1.00
7	3.00	58.00	2.00	10.00	0.23	0.97	0.84	0.14	0.80	0.33	0.54
8	3.00	63.00	1.00	14.00	0.18	0.98	0.81	0.09	0.84	0.29	0.31
9	0.00	82.00	0.00	8.00	0.00	1.00	0.91	1.00	1.00	NaN	0.13
10	0.00	227.00	18.00	16.00	0.00	0.93	0.87	1.00	1.08	NaN	0.13
11	4.00	162.00	0.00	37.00	0.10	1.00	0.82	0.00	0.90	0.18	0.26
12	7.00	178.00	4.00	26.00	0.21	0.98	0.86	0.10	0.81	0.32	0.40
13	4.00	117.00	1.00	18.00	0.18	0.99	0.86	0.05	0.83	0.30	0.25
14	4.00	135.00	1.00	21.00	0.16	0.99	0.86	0.05	0.85	0.27	0.22
15	60.00	195.00	3.00	19.00	0.76	0.98	0.92	0.02	0.24	0.85	1.00
16	3.00	132.00	4.00	26.00	0.10	0.97	0.82	0.28	0.92	0.17	0.86
17	1.00	169.00	0.00	25.00	0.04	1.00	0.87	0.00	0.96	0.07	0.15
18	49.00	178.00	5.00	18.00	0.73	0.97	0.91	0.04	0.28	0.81	1.00
19	40.00	141.00	3.00	10.00	0.80	0.98	0.93	0.03	0.20	0.86	1.00
20	37.00	90.00	4.00	5.00	0.88	0.96	0.93	0.05	0.12	0.89	1.00
21	3.00	126.00	2.00	48.00	0.06	0.98	0.72	0.27	0.96	0.11	0.06
22	24.00	536.00	2.00	70.00	0.26	1.00	0.89	0.01	0.75	0.40	0.20
23	23.00	265.00	1.00	47.00	0.33	1.00	0.86	0.01	0.67	0.49	0.45
24	5.00	282.00	3.00	42.00	0.11	0.99	0.86	0.10	0.90	0.18	0.61
25	16.00	199.00	5.00	50.00	0.24	0.98	0.80	0.10	0.78	0.37	0.20

26	33.00	203.00	0.00	23.00	0.59	1.00	0.91	0.00	0.41	0.74	0.86
27	1.00	293.00	23.00	38.00	0.03	0.93	0.83	2.84	1.05	0.03	0.15
28	4.00	385.00	1.00	53.00	0.07	1.00	0.88	0.04	0.93	0.13	0.16
29	26.00	277.00	9.00	36.00	0.42	0.97	0.87	0.08	0.60	0.54	0.80
30	221.00	132.00	43.00	4.00	0.98	0.75	0.88	0.25	0.02	0.90	0.98
31	8.00	316.00	3.00	99.00	0.07	0.99	0.76	0.13	0.93	0.14	0.07
32	2.00	268.00	3.00	50.00	0.04	0.99	0.84	0.29	0.97	0.07	0.37
33	7.00	190.00	1.00	27.00	0.21	0.99	0.88	0.03	0.80	0.33	0.67
34	2.00	254.00	1.00	37.00	0.05	1.00	0.87	0.08	0.95	0.10	0.16
35	1.00	341.00	4.00	46.00	0.02	0.99	0.87	0.54	0.99	0.04	0.11
36	8.00	195.00	4.00	37.00	0.18	0.98	0.83	0.11	0.84	0.28	0.12
37	1.00	239.00	0.00	36.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
38	3.00	80.00	1.00	12.00	0.20	0.99	0.86	0.06	0.81	0.32	0.31
39	34.00	219.00	2.00	24.00	0.59	0.99	0.91	0.02	0.42	0.72	0.65
40	75.00	136.00	9.00	5.00	0.94	0.94	0.94	0.07	0.07	0.91	1.00
41	4.00	244.00	2.00	46.00	0.08	0.99	0.84	0.10	0.93	0.14	0.43
42	4.00	116.00	0.00	26.00	0.13	1.00	0.82	0.00	0.87	0.24	0.29
43	3.00	241.00	3.00	33.00	0.08	0.99	0.87	0.15	0.93	0.14	0.15
44	6.00	231.00	0.00	36.00	0.14	1.00	0.87	0.00	0.86	0.25	0.15
45	5.00	76.00	2.00	15.00	0.25	0.97	0.83	0.10	0.77	0.37	0.56
46	2.00	282.00	3.00	54.00	0.04	0.99	0.83	0.29	0.97	0.07	0.32
47	31.00	181.00	7.00	29.00	0.52	0.96	0.85	0.07	0.50	0.63	0.51
48	11.00	229.00	6.00	56.00	0.16	0.97	0.79	0.16	0.86	0.26	0.17
49	27.00	140.00	0.00	8.00	0.77	1.00	0.95	0.00	0.23	0.87	1.00
50	8.00	121.00	1.00	44.00	0.15	0.99	0.74	0.05	0.85	0.26	0.08
51	13.00	232.00	4.00	33.00	0.28	0.98	0.87	0.06	0.73	0.41	0.62
52	12.00	94.00	0.00	9.00	0.57	1.00	0.92	0.00	0.43	0.73	1.00
53	3.00	223.00	1.00	51.00	0.06	1.00	0.81	0.08	0.95	0.10	0.17
54	47.00	179.00	5.00	20.00	0.70	0.97	0.90	0.04	0.31	0.79	0.51
55	78.00	273.00	1.00	16.00	0.83	1.00	0.95	0.00	0.17	0.90	1.00
56	14.00	123.00	5.00	26.00	0.35	0.96	0.82	0.11	0.68	0.47	0.45
57	37.00	250.00	0.00	28.00	0.57	1.00	0.91	0.00	0.43	0.73	0.89

58	13.00	200.00	3.00	24.00	0.35	0.99	0.89	0.04	0.66	0.49	0.82
59	40.00	141.00	3.00	10.00	0.80	0.98	0.93	0.03	0.20	0.86	1.00
60	37.00	90.00	4.00	5.00	0.88	0.96	0.93	0.05	0.12	0.89	1.00

Table 6: Statistical Results for Speech Tracking Using Relative Threshold for DTW

Effects of Relative Threshold on Performance											
DTW+MFCC											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	22.00	52.00	3.00	3.00	0.88	0.95	0.93	0.06	0.13	0.88	1.00
2	17.00	38.00	1.00	1.00	0.94	0.97	0.96	0.03	0.06	0.94	1.00
3	53.00	74.00	2.00	3.00	0.95	0.97	0.96	0.03	0.06	0.95	1.00
4	33.00	58.00	5.00	4.00	0.89	0.92	0.91	0.09	0.12	0.88	1.00
5	40.00	53.00	7.00	2.00	0.95	0.88	0.91	0.12	0.05	0.90	1.00
6	44.00	46.00	5.00	0.00	1.00	0.90	0.95	0.10	0.00	0.95	1.00
7	19.00	48.00	6.00	3.00	0.86	0.89	0.88	0.13	0.15	0.81	1.00
8	22.00	66.00	4.00	1.00	0.96	0.94	0.95	0.06	0.05	0.90	1.00
9	45.00	50.00	5.00	0.00	1.00	0.91	0.95	0.09	0.00	0.95	1.00
10	142.00	148.00	19.00	4.00	0.97	0.89	0.93	0.12	0.03	0.93	1.00
11	99.00	129.00	9.00	0.00	1.00	0.93	0.96	0.07	0.00	0.96	1.00
12	118.00	71.00	26.00	1.00	0.99	0.73	0.88	0.27	0.01	0.90	1.00
13	52.00	95.00	4.00	3.00	0.95	0.96	0.95	0.04	0.06	0.94	1.00
14	57.00	105.00	5.00	3.00	0.95	0.95	0.95	0.05	0.05	0.93	1.00
15	121.00	142.00	13.00	1.00	0.99	0.92	0.95	0.08	0.01	0.95	1.00
16	22.00	106.00	20.00	12.00	0.65	0.84	0.80	0.25	0.42	0.58	0.83
17	71.00	115.00	11.00	1.00	0.99	0.91	0.94	0.09	0.02	0.92	1.00
18	105.00	106.00	33.00	9.00	0.92	0.76	0.83	0.26	0.10	0.83	0.93
19	66.00	105.00	22.00	7.00	0.90	0.83	0.85	0.19	0.12	0.82	1.00
20	60.00	59.00	13.00	2.00	0.97	0.82	0.89	0.19	0.04	0.89	1.00
21	129.00	154.00	21.00	2.00	0.98	0.88	0.92	0.12	0.02	0.92	1.00
22	277.00	298.00	33.00	3.00	0.99	0.90	0.94	0.10	0.01	0.94	1.00
23	158.00	199.00	19.00	3.00	0.98	0.91	0.94	0.09	0.02	0.93	1.00
24	143.00	184.00	24.00	2.00	0.99	0.88	0.93	0.12	0.02	0.92	1.00

25	54.00	201.00	32.00	18.00	0.75	0.86	0.84	0.18	0.29	0.68	0.63
26	98.00	153.00	7.00	5.00	0.95	0.96	0.95	0.05	0.05	0.94	1.00
27	230.00	22.00	117.00	7.00	0.97	0.16	0.67	0.87	0.19	0.79	0.97
28	62.00	283.00	51.00	27.00	0.70	0.85	0.82	0.22	0.36	0.61	0.81
29	141.00	205.00	13.00	5.00	0.97	0.94	0.95	0.06	0.04	0.94	1.00
30	231.00	95.00	74.00	2.00	0.99	0.56	0.81	0.44	0.02	0.86	1.00
31	53.00	303.00	166.00	29.00	0.65	0.65	0.65	0.55	0.55	0.35	0.59
32	98.00	209.00	10.00	8.00	0.92	0.95	0.94	0.05	0.08	0.92	1.00
33	53.00	162.00	4.00	12.00	0.82	0.98	0.93	0.03	0.19	0.87	1.00
34	103.00	164.00	21.00	7.00	0.94	0.89	0.91	0.12	0.07	0.88	1.00
35	18.00	242.00	76.00	40.00	0.31	0.76	0.69	0.77	0.91	0.24	0.54
36	40.00	179.00	6.00	15.00	0.73	0.97	0.91	0.04	0.28	0.79	0.63
37	96.00	185.00	3.00	3.00	0.97	0.98	0.98	0.02	0.03	0.97	1.00
38	37.00	58.00	3.00	0.00	1.00	0.95	0.97	0.05	0.00	0.96	1.00
39	159.00	110.00	22.00	0.00	1.00	0.83	0.92	0.17	0.00	0.94	1.00
40	105.00	104.00	16.00	3.00	0.97	0.87	0.92	0.14	0.03	0.92	1.00
41	80.00	200.00	6.00	4.00	0.95	0.97	0.97	0.03	0.05	0.94	1.00
42	84.00	75.00	10.00	1.00	0.99	0.88	0.94	0.12	0.01	0.94	1.00
43	122.00	157.00	8.00	3.00	0.98	0.95	0.96	0.05	0.03	0.96	1.00
44	35.00	157.00	48.00	25.00	0.58	0.77	0.72	0.40	0.54	0.49	0.88
45	41.00	55.00	9.00	2.00	0.95	0.86	0.90	0.15	0.05	0.88	1.00
46	81.00	243.00	5.00	14.00	0.85	0.98	0.94	0.02	0.15	0.90	0.89
47	96.00	124.00	38.00	9.00	0.91	0.77	0.82	0.26	0.11	0.80	0.90
48	52.00	95.00	4.00	3.00	0.95	0.96	0.95	0.04	0.06	0.94	1.00
49	57.00	105.00	5.00	3.00	0.95	0.95	0.95	0.05	0.05	0.93	1.00
50	121.00	142.00	13.00	1.00	0.99	0.92	0.95	0.08	0.01	0.95	1.00
51	22.00	106.00	20.00	12.00	0.65	0.84	0.80	0.25	0.42	0.58	0.83
52	71.00	115.00	11.00	1.00	0.99	0.91	0.94	0.09	0.02	0.92	1.00
53	105.00	106.00	33.00	9.00	0.92	0.76	0.83	0.26	0.10	0.83	0.93
54	35.00	157.00	48.00	25.00	0.58	0.77	0.72	0.40	0.54	0.49	0.88
55	41.00	55.00	9.00	2.00	0.95	0.86	0.90	0.15	0.05	0.88	1.00
56	81.00	243.00	5.00	14.00	0.85	0.98	0.94	0.02	0.15	0.90	0.89

57	96.00	124.00	38.00	9.00	0.91	0.77	0.82	0.26	0.11	0.80	0.90
58	52.00	95.00	4.00	3.00	0.95	0.96	0.95	0.04	0.06	0.94	1.00
59	66.00	105.00	22.00	7.00	0.90	0.83	0.85	0.19	0.12	0.82	1.00
60	60.00	59.00	13.00	2.00	0.97	0.82	0.89	0.19	0.04	0.89	1.00

Table 7: Statistical Results for Non-Adaptive Speech Tracking Based on DTW and MFCC

Search Region Based Non-Adaptive Tracking & Similarity Measurement (DTW+MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	4.00	81.00	1.00	12.00	0.25	0.99	0.87	0.05	0.76	0.38	0.21
2	5.00	52.00	5.00	8.00	0.38	0.91	0.81	0.23	0.67	0.43	0.30
3	55.00	87.00	12.00	0.00	1.00	0.88	0.92	0.12	0.00	0.90	1.00
4	41.00	66.00	12.00	0.00	1.00	0.85	0.90	0.15	0.00	0.87	1.00
5	19.00	90.00	7.00	10.00	0.66	0.93	0.87	0.11	0.37	0.69	0.50
6	31.00	67.00	7.00	0.00	1.00	0.91	0.93	0.09	0.00	0.90	1.00
7	26.00	53.00	5.00	0.00	1.00	0.91	0.94	0.09	0.00	0.91	1.00
8	10.00	88.00	3.00	11.00	0.48	0.97	0.88	0.07	0.54	0.59	0.44
9	39.00	69.00	4.00	0.00	1.00	0.95	0.96	0.05	0.00	0.95	1.00
10	55.00	265.00	7.00	30.00	0.65	0.97	0.90	0.04	0.36	0.75	0.43
11	85.00	164.00	17.00	0.00	1.00	0.91	0.94	0.09	0.00	0.91	1.00
12	86.00	142.00	17.00	0.00	1.00	0.89	0.93	0.11	0.00	0.91	1.00
13	52.00	104.00	11.00	1.00	0.98	0.90	0.93	0.10	0.02	0.90	1.00
14	48.00	132.00	8.00	1.00	0.98	0.94	0.95	0.06	0.02	0.91	1.00
15	73.00	215.00	14.00	13.00	0.85	0.94	0.91	0.07	0.16	0.84	0.73
16	9.00	161.00	3.00	23.00	0.28	0.98	0.87	0.07	0.73	0.41	0.25
17	68.00	139.00	16.00	1.00	0.99	0.90	0.92	0.10	0.02	0.89	1.00
18	35.00	229.00	4.00	26.00	0.57	0.98	0.90	0.03	0.43	0.70	0.40
19	75.00	127.00	22.00	0.00	1.00	0.85	0.90	0.15	0.00	0.87	1.00
20	42.00	98.00	12.00	2.00	0.95	0.89	0.91	0.11	0.05	0.86	1.00
21	117.00	215.00	18.00	0.00	1.00	0.92	0.95	0.08	0.00	0.93	1.00
22	222.00	440.00	41.00	4.00	0.98	0.91	0.94	0.09	0.02	0.91	1.00
23	120.00	287.00	16.00	4.00	0.97	0.95	0.95	0.05	0.03	0.92	1.00

24	52.00	313.00	10.00	31.00	0.63	0.97	0.90	0.05	0.39	0.72	0.48
25	129.00	242.00	35.00	0.00	1.00	0.87	0.91	0.13	0.00	0.88	1.00
26	94.00	195.00	11.00	1.00	0.99	0.95	0.96	0.05	0.01	0.94	1.00
27	3.00	353.00	18.00	60.00	0.05	0.95	0.82	1.02	1.00	0.07	0.05
28	161.00	344.00	26.00	1.00	0.99	0.93	0.95	0.07	0.01	0.92	1.00
29	33.00	327.00	21.00	46.00	0.42	0.94	0.84	0.14	0.62	0.50	0.26
30	83.00	335.00	15.00	29.00	0.74	0.96	0.90	0.06	0.27	0.79	0.58
31	141.00	492.00	50.00	38.00	0.79	0.91	0.88	0.12	0.23	0.76	0.64
32	29.00	300.00	6.00	43.00	0.40	0.98	0.87	0.05	0.61	0.54	0.22
33	4.00	216.00	4.00	35.00	0.10	0.98	0.85	0.18	0.91	0.17	0.08
34	17.00	287.00	3.00	43.00	0.28	0.99	0.87	0.04	0.72	0.42	0.16
35	65.00	364.00	10.00	30.00	0.68	0.97	0.91	0.04	0.32	0.76	0.57
36	71.00	208.00	5.00	3.00	0.96	0.98	0.97	0.02	0.04	0.95	1.00
37	35.00	263.00	1.00	30.00	0.54	1.00	0.91	0.01	0.46	0.69	0.38
38	36.00	71.00	4.00	1.00	0.97	0.95	0.96	0.05	0.03	0.94	1.00
39	109.00	201.00	12.00	0.00	1.00	0.94	0.96	0.06	0.00	0.95	1.00
40	9.00	224.00	0.00	33.00	0.21	1.00	0.88	0.00	0.79	0.35	0.18
41	105.00	226.00	12.00	0.00	1.00	0.95	0.97	0.05	0.00	0.95	1.00
42	60.00	120.00	9.00	0.00	1.00	0.93	0.95	0.07	0.00	0.93	1.00
43	107.00	207.00	14.00	1.00	0.99	0.94	0.95	0.06	0.01	0.93	1.00
44	50.00	251.00	6.00	22.00	0.69	0.98	0.91	0.03	0.31	0.78	0.57
45	38.00	73.00	8.00	0.00	1.00	0.90	0.93	0.10	0.00	0.90	1.00
46	23.00	335.00	2.00	46.00	0.33	0.99	0.88	0.02	0.67	0.49	0.22
47	94.00	199.00	15.00	0.00	1.00	0.93	0.95	0.07	0.00	0.93	1.00
48	133.00	259.00	20.00	1.00	0.99	0.93	0.95	0.07	0.01	0.93	1.00
49	26.00	153.00	7.00	17.00	0.60	0.96	0.88	0.07	0.41	0.68	0.45
50	42.00	258.00	7.00	29.00	0.59	0.97	0.89	0.04	0.42	0.70	0.44
51	94.00	213.00	15.00	0.00	1.00	0.93	0.95	0.07	0.00	0.93	1.00
52	20.00	101.00	1.00	11.00	0.65	0.99	0.91	0.02	0.36	0.77	0.47
53	16.00	308.00	8.00	46.00	0.26	0.97	0.86	0.10	0.76	0.37	0.17
54	93.00	188.00	10.00	3.00	0.97	0.95	0.96	0.05	0.03	0.93	0.95
55	115.00	281.00	22.00	9.00	0.93	0.93	0.93	0.08	0.08	0.88	0.87

56	68.00	146.00	10.00	0.00	1.00	0.94	0.96	0.06	0.00	0.93	1.00
57	33.00	302.00	2.00	41.00	0.45	0.99	0.89	0.01	0.56	0.61	0.26
58	23.00	230.00	9.00	32.00	0.42	0.96	0.86	0.09	0.60	0.53	0.26
59	75.00	127.00	22.00	0.00	1.00	0.85	0.90	0.15	0.00	0.87	1.00
60	42.00	98.00	12.00	2.00	0.95	0.89	0.91	0.11	0.05	0.86	1.00

Table 8: Statistical Results for Speech Tracking Based on DTW and Energy Based Silence Removal

Kalman Filter with Energy & Spectral Centroid Silence Removal Approach (DTW+MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	20.00	53.00	5.00	3.00	0.87	0.91	0.90	0.10	0.14	0.83	1.00
2	22.00	50.00	3.00	1.00	0.96	0.94	0.95	0.06	0.05	0.92	1.00
3	105.00	65.00	36.00	0.00	1.00	0.64	0.83	0.36	0.00	0.85	1.00
4	85.00	73.00	7.00	1.00	0.99	0.91	0.95	0.09	0.01	0.96	1.00
5	64.00	99.00	12.00	1.00	0.98	0.89	0.93	0.11	0.02	0.91	1.00
6	51.00	72.00	21.00	4.00	0.93	0.77	0.83	0.24	0.09	0.80	0.96
7	7.00	67.00	28.00	17.00	0.29	0.71	0.62	1.01	1.00	0.24	0.55
8	71.00	56.00	11.00	0.00	1.00	0.84	0.92	0.16	0.00	0.93	1.00
9	35.00	77.00	8.00	4.00	0.90	0.91	0.90	0.10	0.11	0.85	1.00
10	214.00	138.00	42.00	3.00	0.99	0.77	0.89	0.24	0.02	0.90	1.00
11	265.00	70.00	106.00	5.00	0.98	0.40	0.75	0.61	0.05	0.83	0.97
12	164.00	161.00	18.00	3.00	0.98	0.90	0.94	0.10	0.02	0.94	1.00
13	69.00	107.00	28.00	8.00	0.90	0.79	0.83	0.23	0.13	0.79	0.97
14	97.00	155.00	11.00	3.00	0.97	0.93	0.95	0.07	0.03	0.93	1.00
15	129.00	207.00	32.00	6.00	0.96	0.87	0.90	0.14	0.05	0.87	1.00
16	55.00	112.00	17.00	7.00	0.89	0.87	0.87	0.15	0.13	0.82	1.00
17	122.00	206.00	6.00	4.00	0.97	0.97	0.97	0.03	0.03	0.96	1.00
18	164.00	172.00	52.00	13.00	0.93	0.77	0.84	0.25	0.10	0.83	1.00
19	165.00	128.00	34.00	5.00	0.97	0.79	0.88	0.22	0.04	0.89	1.00
20	92.00	172.00	17.00	6.00	0.94	0.91	0.92	0.10	0.07	0.89	1.00
21	209.00	244.00	35.00	6.00	0.97	0.87	0.92	0.13	0.03	0.91	1.00
22	475.00	356.00	90.00	4.00	0.99	0.80	0.90	0.20	0.01	0.91	1.00

23	329.00	222.00	83.00	6.00	0.98	0.73	0.86	0.28	0.02	0.88	1.00
24	224.00	317.00	113.00	22.00	0.91	0.74	0.80	0.29	0.12	0.77	0.91
25	221.00	325.00	46.00	12.00	0.95	0.88	0.90	0.13	0.06	0.88	1.00
26	219.00	194.00	44.00	7.00	0.97	0.82	0.89	0.19	0.04	0.90	1.00
27	283.00	230.00	70.00	9.00	0.97	0.77	0.87	0.24	0.04	0.88	0.98
28	356.00	166.00	109.00	5.00	0.99	0.60	0.82	0.40	0.02	0.86	1.00
29	248.00	326.00	100.00	21.00	0.92	0.77	0.83	0.25	0.10	0.80	1.00
30	319.00	221.00	92.00	10.00	0.97	0.71	0.84	0.30	0.04	0.86	1.00
31	482.00	608.00	54.00	12.00	0.98	0.92	0.94	0.08	0.03	0.94	1.00
32	309.00	110.00	86.00	2.00	0.99	0.56	0.83	0.44	0.01	0.88	1.00
33	139.00	185.00	20.00	6.00	0.96	0.90	0.93	0.10	0.05	0.91	1.00
34	221.00	156.00	58.00	3.00	0.99	0.73	0.86	0.27	0.02	0.88	1.00
35	360.00	168.00	99.00	2.00	0.99	0.63	0.84	0.37	0.01	0.88	1.00
36	153.00	207.00	30.00	9.00	0.94	0.87	0.90	0.13	0.06	0.89	1.00
37	174.00	276.00	19.00	7.00	0.96	0.94	0.95	0.07	0.04	0.93	1.00
38	63.00	108.00	12.00	5.00	0.93	0.90	0.91	0.11	0.08	0.88	1.00
39	205.00	123.00	62.00	5.00	0.98	0.66	0.83	0.34	0.04	0.86	1.00
40	282.00	107.00	54.00	2.00	0.99	0.66	0.87	0.34	0.01	0.91	1.00
41	64.00	99.00	12.00	1.00	0.98	0.89	0.93	0.11	0.02	0.91	1.00
42	51.00	72.00	21.00	4.00	0.93	0.77	0.83	0.24	0.09	0.80	0.96
43	7.00	67.00	28.00	17.00	0.29	0.71	0.62	1.01	1.00	0.24	0.55
44	71.00	56.00	11.00	0.00	1.00	0.84	0.92	0.16	0.00	0.93	1.00
45	35.00	77.00	8.00	4.00	0.90	0.91	0.90	0.10	0.11	0.85	1.00
46	214.00	138.00	42.00	3.00	0.99	0.77	0.89	0.24	0.02	0.90	1.00
47	265.00	70.00	106.00	5.00	0.98	0.40	0.75	0.61	0.05	0.83	0.97
48	164.00	161.00	18.00	3.00	0.98	0.90	0.94	0.10	0.02	0.94	1.00
49	69.00	107.00	28.00	8.00	0.90	0.79	0.83	0.23	0.13	0.79	0.97
50	153.00	207.00	30.00	9.00	0.94	0.87	0.90	0.13	0.06	0.89	1.00
51	174.00	276.00	19.00	7.00	0.96	0.94	0.95	0.07	0.04	0.93	1.00
52	63.00	108.00	12.00	5.00	0.93	0.90	0.91	0.11	0.08	0.88	1.00
53	205.00	123.00	62.00	5.00	0.98	0.66	0.83	0.34	0.04	0.86	1.00
54	282.00	107.00	54.00	2.00	0.99	0.66	0.87	0.34	0.01	0.91	1.00

55	153.00	207.00	30.00	9.00	0.94	0.87	0.90	0.13	0.06	0.89	1.00
56	174.00	276.00	19.00	7.00	0.96	0.94	0.95	0.07	0.04	0.93	1.00
57	63.00	108.00	12.00	5.00	0.93	0.90	0.91	0.11	0.08	0.88	1.00
58	205.00	123.00	62.00	5.00	0.98	0.66	0.83	0.34	0.04	0.86	1.00
59	165.00	128.00	34.00	5.00	0.97	0.79	0.88	0.22	0.04	0.89	1.00
60	92.00	172.00	17.00	6.00	0.94	0.91	0.92	0.10	0.07	0.89	1.00

Table 9: Statistical Results for Speech Tracking Based on Constrained DTW and MFCC

Kalman Filter Based Adaptive Speech Tracking & Similarity Measurement (DTW Constrained + MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	1.00	65.00	0.00	9.00	0.10	1.00	0.88	0.00	0.90	0.18	0.40
2	1.00	51.00	0.00	7.00	0.13	1.00	0.88	0.00	0.88	0.22	0.50
3	0.00	104.00	0.00	10.00	0.00	1.00	0.91	1.00	1.00	NaN	0.10
4	0.00	82.00	0.00	8.00	0.00	1.00	0.91	1.00	1.00	NaN	0.13
5	0.00	82.00	0.00	8.00	0.00	1.00	0.91	1.00	1.00	NaN	0.13
6	0.00	71.00	0.00	7.00	0.00	1.00	0.91	1.00	1.00	NaN	0.14
7	1.00	65.00	0.00	9.00	0.10	1.00	0.88	0.00	0.90	0.18	0.40
8	0.00	71.00	0.00	7.00	0.00	1.00	0.91	1.00	1.00	NaN	0.14
9	0.00	82.00	0.00	8.00	0.00	1.00	0.91	1.00	1.00	NaN	0.13
10	0.00	258.00	0.00	24.00	0.00	1.00	0.91	1.00	1.00	NaN	0.04
11	1.00	194.00	0.00	29.00	0.03	1.00	0.87	0.00	0.97	0.06	0.13
12	1.00	181.00	0.00	27.00	0.04	1.00	0.87	0.00	0.96	0.07	0.14
13	0.00	115.00	0.00	11.00	0.00	1.00	0.91	1.00	1.00	NaN	0.09
14	0.00	137.00	0.00	13.00	0.00	1.00	0.91	1.00	1.00	NaN	0.08
15	1.00	233.00	0.00	35.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
16	1.00	142.00	0.00	21.00	0.05	1.00	0.87	0.00	0.95	0.09	0.18
17	1.00	169.00	0.00	25.00	0.04	1.00	0.87	0.00	0.96	0.07	0.15
18	0.00	203.00	0.00	19.00	0.00	1.00	0.91	1.00	1.00	NaN	0.05
19	1.00	169.00	0.00	25.00	0.04	1.00	0.87	0.00	0.96	0.07	0.15
20	0.00	104.00	0.00	10.00	0.00	1.00	0.91	1.00	1.00	NaN	0.10
21	1.00	260.00	0.00	39.00	0.03	1.00	0.87	0.00	0.97	0.05	0.10

22	20.00	496.00	1.00	69.00	0.22	1.00	0.88	0.01	0.78	0.36	0.44
23	2.00	256.00	0.00	60.00	0.03	1.00	0.81	0.00	0.97	0.06	0.16
24	0.00	291.00	0.00	27.00	0.00	1.00	0.92	1.00	1.00	NaN	0.04
25	1.00	302.00	0.00	45.00	0.02	1.00	0.87	0.00	0.98	0.04	0.09
26	1.00	221.00	0.00	33.00	0.03	1.00	0.87	0.00	0.97	0.06	0.12
27	1.00	323.00	0.00	48.00	0.02	1.00	0.87	0.00	0.98	0.04	0.08
28	1.00	407.00	0.00	60.00	0.02	1.00	0.87	0.00	0.98	0.03	0.07
29	1.00	316.00	0.00	47.00	0.02	1.00	0.87	0.00	0.98	0.04	0.08
30	1.00	344.00	0.00	51.00	0.02	1.00	0.87	0.00	0.98	0.04	0.08
31	1.00	554.00	0.00	81.00	0.01	1.00	0.87	0.00	0.99	0.02	0.05
32	1.00	274.00	0.00	41.00	0.02	1.00	0.87	0.00	0.98	0.05	0.10
33	1.00	187.00	0.00	28.00	0.03	1.00	0.87	0.00	0.97	0.07	0.14
34	1.00	260.00	0.00	39.00	0.03	1.00	0.87	0.00	0.97	0.05	0.10
35	0.00	335.00	0.00	31.00	0.00	1.00	0.92	1.00	1.00	NaN	0.03
36	1.00	215.00	0.00	32.00	0.03	1.00	0.87	0.00	0.97	0.06	0.12
37	1.00	239.00	0.00	36.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
38	1.00	83.00	0.00	12.00	0.08	1.00	0.88	0.00	0.92	0.14	0.31
39	3.00	195.00	0.00	44.00	0.06	1.00	0.82	0.00	0.94	0.12	0.19
40	3.00	160.00	0.00	35.00	0.08	1.00	0.82	0.00	0.92	0.15	0.21
41	1.00	253.00	0.00	38.00	0.03	1.00	0.87	0.00	0.97	0.05	0.10
42	1.00	142.00	0.00	21.00	0.05	1.00	0.87	0.00	0.95	0.09	0.18
43	1.00	239.00	0.00	36.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
44	1.00	239.00	0.00	36.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
45	1.00	90.00	0.00	13.00	0.07	1.00	0.88	0.00	0.93	0.13	0.29
46	1.00	302.00	0.00	45.00	0.02	1.00	0.87	0.00	0.98	0.04	0.09
47	2.00	151.00	1.00	43.00	0.04	0.99	0.78	0.15	0.96	0.08	0.11
48	1.00	309.00	0.00	46.00	0.02	1.00	0.87	0.00	0.98	0.04	0.09
49	6.00	136.00	1.00	22.00	0.21	0.99	0.86	0.03	0.79	0.34	1.00
50	1.00	246.00	0.00	37.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
51	1.00	239.00	0.00	36.00	0.03	1.00	0.87	0.00	0.97	0.05	0.11
52	2.00	84.00	0.00	17.00	0.11	1.00	0.83	0.00	0.89	0.19	0.53
53	0.00	269.00	0.00	25.00	0.00	1.00	0.91	1.00	1.00	NaN	0.04

54	3.00	177.00	0.00	39.00	0.07	1.00	0.82	0.00	0.93	0.13	0.24
55	1.00	316.00	0.00	47.00	0.02	1.00	0.87	0.00	0.98	0.04	0.08
56	5.00	113.00	1.00	30.00	0.14	0.99	0.79	0.06	0.86	0.24	0.15
57	1.00	274.00	0.00	41.00	0.02	1.00	0.87	0.00	0.98	0.05	0.10
58	1.00	215.00	0.00	32.00	0.03	1.00	0.87	0.00	0.97	0.06	0.12
59	1.00	169.00	0.00	25.00	0.04	1.00	0.87	0.00	0.96	0.07	0.15
60	0.00	104.00	0.00	10.00	0.00	1.00	0.91	1.00	1.00	NaN	0.10

Table 10: Statistical Results for Speech Tracking Based Relative Threshold for Constrained DTW

Effects of Relative Threshold on Performance DTW Constrained + MFCC											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	12.00	69.00	0.00	4.00	0.75	1.00	0.95	0.00	0.25	0.86	1.00
2	12.00	39.00	1.00	3.00	0.80	0.97	0.93	0.03	0.21	0.86	1.00
3	30.00	96.00	1.00	5.00	0.86	0.99	0.95	0.01	0.14	0.91	1.00
4	25.00	78.00	0.00	4.00	0.86	1.00	0.96	0.00	0.14	0.93	1.00
5	15.00	85.00	0.00	6.00	0.71	1.00	0.94	0.00	0.29	0.83	1.00
6	22.00	69.00	1.00	2.00	0.92	0.99	0.97	0.02	0.08	0.94	1.00
7	16.00	59.00	1.00	4.00	0.80	0.98	0.94	0.02	0.20	0.86	1.00
8	9.00	81.00	0.00	7.00	0.56	1.00	0.93	0.00	0.44	0.72	1.00
9	16.00	81.00	0.00	3.00	0.84	1.00	0.97	0.00	0.16	0.91	1.00
10	0.00	244.00	9.00	22.00	0.00	0.96	0.89	Inf	1.04	NaN	0.09
11	63.00	163.00	3.00	3.00	0.95	0.98	0.97	0.02	0.05	0.95	1.00
12	43.00	167.00	0.00	7.00	0.86	1.00	0.97	0.00	0.14	0.92	1.00
13	30.00	119.00	0.00	4.00	0.88	1.00	0.97	0.00	0.12	0.94	1.00
14	3.00	140.00	0.00	21.00	0.13	1.00	0.87	0.00	0.88	0.22	0.18
15	14.00	196.00	0.00	34.00	0.29	1.00	0.86	0.00	0.71	0.45	0.42
16	1.00	142.00	0.00	21.00	0.05	1.00	0.87	0.00	0.95	0.09	0.18
17	36.00	152.00	1.00	6.00	0.86	0.99	0.96	0.01	0.14	0.91	1.00
18	41.00	205.00	1.00	9.00	0.82	1.00	0.96	0.01	0.18	0.89	1.00
19	45.00	151.00	1.00	5.00	0.90	0.99	0.97	0.01	0.10	0.94	1.00
20	23.00	106.00	0.00	6.00	0.79	1.00	0.96	0.00	0.21	0.88	1.00

21	82.00	226.00	1.00	3.00	0.96	1.00	0.99	0.00	0.04	0.98	1.00
22	115.00	479.00	4.00	16.00	0.88	0.99	0.97	0.01	0.12	0.92	1.00
23	78.00	286.00	3.00	6.00	0.93	0.99	0.98	0.01	0.07	0.95	1.00
24	67.00	270.00	3.00	9.00	0.88	0.99	0.97	0.01	0.12	0.92	1.00
25	84.00	243.00	7.00	17.00	0.83	0.97	0.93	0.03	0.17	0.87	1.00
26	50.00	206.00	1.00	7.00	0.88	1.00	0.97	0.01	0.12	0.93	1.00
27	58.00	304.00	1.00	17.00	0.77	1.00	0.95	0.00	0.23	0.87	1.00
28	83.00	365.00	1.00	16.00	0.84	1.00	0.96	0.00	0.16	0.91	1.00
29	73.00	286.00	0.00	12.00	0.86	1.00	0.97	0.00	0.14	0.92	1.00
30	84.00	305.00	0.00	8.00	0.91	1.00	0.98	0.00	0.09	0.95	1.00
31	33.00	429.00	3.00	83.00	0.28	0.99	0.84	0.02	0.72	0.43	0.29
32	40.00	263.00	2.00	21.00	0.66	0.99	0.93	0.01	0.35	0.78	1.00
33	8.00	171.00	0.00	29.00	0.22	1.00	0.86	0.00	0.78	0.36	0.54
34	14.00	215.00	0.00	40.00	0.26	1.00	0.85	0.00	0.74	0.41	0.38
35	83.00	307.00	5.00	8.00	0.91	0.98	0.97	0.02	0.09	0.93	1.00
36	45.00	203.00	0.00	11.00	0.80	1.00	0.96	0.00	0.20	0.89	1.00
37	4.00	152.00	1.00	44.00	0.08	0.99	0.78	0.08	0.92	0.15	0.11
38	15.00	81.00	0.00	5.00	0.75	1.00	0.95	0.00	0.25	0.86	1.00
39	50.00	224.00	0.00	9.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00
40	60.00	163.00	1.00	2.00	0.97	0.99	0.99	0.01	0.03	0.98	1.00
41	46.00	246.00	0.00	9.00	0.84	1.00	0.97	0.00	0.16	0.91	1.00
42	31.00	136.00	1.00	4.00	0.89	0.99	0.97	0.01	0.12	0.93	1.00
43	39.00	238.00	0.00	11.00	0.78	1.00	0.96	0.00	0.22	0.88	1.00
44	21.00	219.00	3.00	31.00	0.40	0.99	0.88	0.03	0.60	0.55	0.51
45	19.00	92.00	0.00	1.00	0.95	1.00	0.99	0.00	0.05	0.97	1.00
46	57.00	293.00	0.00	6.00	0.90	1.00	0.98	0.00	0.10	0.95	1.00
47	45.00	211.00	1.00	12.00	0.79	1.00	0.95	0.01	0.21	0.87	1.00
48	59.00	286.00	1.00	16.00	0.79	1.00	0.95	0.00	0.21	0.87	1.00
49	28.00	144.00	0.00	5.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00
50	53.00	233.00	0.00	9.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00
51	35.00	235.00	0.00	17.00	0.67	1.00	0.94	0.00	0.33	0.80	1.00
52	17.00	89.00	0.00	3.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00

53	40.00	269.00	0.00	17.00	0.70	1.00	0.95	0.00	0.30	0.82	1.00
54	37.00	214.00	0.00	8.00	0.82	1.00	0.97	0.00	0.18	0.90	1.00
55	59.00	294.00	0.00	12.00	0.83	1.00	0.97	0.00	0.17	0.91	1.00
56	32.00	159.00	0.00	7.00	0.82	1.00	0.96	0.00	0.18	0.90	1.00
57	42.00	270.00	1.00	16.00	0.72	1.00	0.95	0.01	0.28	0.83	1.00
58	53.00	198.00	2.00	5.00	0.91	0.99	0.97	0.01	0.09	0.94	1.00
59	45.00	151.00	1.00	5.00	0.90	0.99	0.97	0.01	0.10	0.94	1.00
60	23.00	106.00	0.00	6.00	0.79	1.00	0.96	0.00	0.21	0.88	1.00

Table 11: Statistical Results for Non-Adaptive Speech Tracking Based on Constrained DTW

Search Region Based Non-Adaptive Tracking & Similarity Measurement (DTW Constrained + MFCC)											
Test Cases	TP	TN	FP	FN	Sen	Spec	Acc	LR+	LR-	F-Score	Tracking Accuracy
1	2.00	83.00	0.00	13.00	0.13	1.00	0.87	0.00	0.87	0.24	0.14
2	4.00	58.00	0.00	8.00	0.33	1.00	0.89	0.00	0.67	0.50	0.30
3	1.00	132.00	0.00	21.00	0.05	1.00	0.86	0.00	0.95	0.09	0.09
4	2.00	108.00	0.00	16.00	0.11	1.00	0.87	0.00	0.89	0.20	0.17
5	1.00	108.00	0.00	17.00	0.06	1.00	0.87	0.00	0.94	0.11	0.11
6	1.00	96.00	0.00	15.00	0.06	1.00	0.87	0.00	0.94	0.12	0.13
7	1.00	78.00	0.00	12.00	0.08	1.00	0.87	0.00	0.92	0.14	0.15
8	1.00	96.00	0.00	15.00	0.06	1.00	0.87	0.00	0.94	0.12	0.13
9	5.00	102.00	0.00	12.00	0.29	1.00	0.90	0.00	0.71	0.45	0.35
10	9.00	303.00	0.00	45.00	0.17	1.00	0.87	0.00	0.83	0.29	0.16
11	7.00	232.00	0.00	34.00	0.17	1.00	0.88	0.00	0.83	0.29	0.18
12	15.00	212.00	0.00	25.00	0.38	1.00	0.90	0.00	0.63	0.55	0.33
13	7.00	154.00	0.00	21.00	0.25	1.00	0.88	0.00	0.75	0.40	0.23
14	3.00	166.00	0.00	27.00	0.10	1.00	0.86	0.00	0.90	0.18	0.07
15	2.00	270.00	0.00	43.00	0.04	1.00	0.86	0.00	0.96	0.09	0.07
16	4.00	168.00	0.00	24.00	0.14	1.00	0.88	0.00	0.86	0.25	0.21
17	20.00	194.00	0.00	17.00	0.54	1.00	0.93	0.00	0.46	0.70	0.52
18	1.00	252.00	0.00	41.00	0.02	1.00	0.86	0.00	0.98	0.05	0.05
19	3.00	197.00	0.00	31.00	0.09	1.00	0.87	0.00	0.91	0.16	0.09

20	10.00	135.00	0.00	16.00	0.38	1.00	0.90	0.00	0.62	0.56	0.35
21	6.00	306.00	0.00	45.00	0.12	1.00	0.87	0.00	0.88	0.21	0.14
22	4.00	608.00	3.00	99.00	0.04	1.00	0.86	0.13	0.97	0.07	0.04
23	3.00	369.00	2.00	60.00	0.05	0.99	0.86	0.11	0.96	0.09	0.05
24	2.00	348.00	0.00	56.00	0.03	1.00	0.86	0.00	0.97	0.07	0.05
25	5.00	354.00	0.00	54.00	0.08	1.00	0.87	0.00	0.92	0.16	0.10
26	5.00	264.00	0.00	39.00	0.11	1.00	0.87	0.00	0.89	0.20	0.14
27	2.00	372.00	0.00	60.00	0.03	1.00	0.86	0.00	0.97	0.06	0.05
28	1.00	462.00	0.00	76.00	0.01	1.00	0.86	0.00	0.99	0.03	0.03
29	2.00	366.00	0.00	59.00	0.03	1.00	0.86	0.00	0.97	0.06	0.05
30	1.00	396.00	0.00	65.00	0.02	1.00	0.86	0.00	0.98	0.03	0.03
31	5.00	618.00	0.00	98.00	0.05	1.00	0.86	0.00	0.95	0.09	0.06
32	1.00	324.00	0.00	53.00	0.02	1.00	0.86	0.00	0.98	0.04	0.04
33	2.00	222.00	0.00	35.00	0.05	1.00	0.86	0.00	0.95	0.10	0.08
34	1.00	300.00	0.00	49.00	0.02	1.00	0.86	0.00	0.98	0.04	0.04
35	1.00	402.00	0.00	66.00	0.01	1.00	0.86	0.00	0.99	0.03	0.03
36	3.00	251.00	0.00	40.00	0.07	1.00	0.86	0.00	0.93	0.13	0.07
37	1.00	282.00	0.00	46.00	0.02	1.00	0.86	0.00	0.98	0.04	0.04
38	8.00	100.00	0.00	11.00	0.42	1.00	0.91	0.00	0.58	0.59	0.41
39	16.00	277.00	0.00	36.00	0.31	1.00	0.89	0.00	0.69	0.47	0.28
40	1.00	228.00	0.00	37.00	0.03	1.00	0.86	0.00	0.97	0.05	0.05
41	8.00	299.00	0.00	43.00	0.16	1.00	0.88	0.00	0.84	0.27	0.18
42	2.00	168.00	0.00	26.00	0.07	1.00	0.87	0.00	0.93	0.13	0.11
43	3.00	288.00	0.00	45.00	0.06	1.00	0.87	0.00	0.94	0.12	0.08
44	14.00	280.00	0.00	35.00	0.29	1.00	0.89	0.00	0.71	0.44	0.28
45	11.00	106.00	0.00	9.00	0.55	1.00	0.93	0.00	0.45	0.71	0.56
46	14.00	345.00	0.00	47.00	0.23	1.00	0.88	0.00	0.77	0.37	0.21
47	10.00	267.00	0.00	38.00	0.21	1.00	0.88	0.00	0.79	0.34	0.18
48	23.00	355.00	0.00	42.00	0.35	1.00	0.90	0.00	0.65	0.52	0.32
49	7.00	173.00	0.00	23.00	0.23	1.00	0.89	0.00	0.77	0.38	0.24
50	6.00	286.00	0.00	44.00	0.12	1.00	0.87	0.00	0.88	0.21	0.10
51	2.00	281.00	0.00	46.00	0.04	1.00	0.86	0.00	0.96	0.08	0.04

52	2.00	114.00	0.00	17.00	0.11	1.00	0.87	0.00	0.89	0.19	0.16
53	5.00	322.00	0.00	51.00	0.09	1.00	0.87	0.00	0.91	0.16	0.07
54	7.00	250.00	1.00	36.00	0.16	1.00	0.87	0.02	0.84	0.27	0.17
55	2.00	366.00	0.00	59.00	0.03	1.00	0.86	0.00	0.97	0.06	0.05
56	4.00	198.00	0.00	29.00	0.12	1.00	0.87	0.00	0.88	0.22	0.15
57	10.00	323.00	0.00	45.00	0.18	1.00	0.88	0.00	0.82	0.31	0.19
58	5.00	251.00	0.00	38.00	0.12	1.00	0.87	0.00	0.88	0.21	0.12
59	3.00	197.00	0.00	31.00	0.09	1.00	0.87	0.00	0.91	0.16	0.09
60	10.00	135.00	0.00	16.00	0.38	1.00	0.90	0.00	0.62	0.56	0.35

Table 12: Statistical Results for Speech Tracking Based Constrained DTW and Traditional Silence Removal

Kalman Filter with Energy & Spectral Centroid Silence Removal Approach (DTW Constrained + MFCC)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	23.00	58.00	2.00	0.00	1.00	0.97	0.98	0.03	0.00	0.96	1.00
2	13.00	59.00	0.00	2.00	0.87	1.00	0.97	0.00	0.13	0.93	1.00
3	38.00	158.00	2.00	4.00	0.90	0.99	0.97	0.01	0.10	0.93	1.00
4	40.00	124.00	1.00	1.00	0.98	0.99	0.99	0.01	0.02	0.98	1.00
5	29.00	138.00	1.00	6.00	0.83	0.99	0.96	0.01	0.17	0.89	1.00
6	25.00	116.00	0.00	5.00	0.83	1.00	0.97	0.00	0.17	0.91	1.00
7	5.00	98.00	2.00	18.00	0.22	0.98	0.84	0.09	0.80	0.33	0.50
8	8.00	75.00	1.00	18.00	0.31	0.99	0.81	0.04	0.70	0.46	0.32
9	6.00	97.00	1.00	17.00	0.26	0.99	0.85	0.04	0.75	0.40	0.38
10	76.00	308.00	0.00	13.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00
11	104.00	330.00	6.00	9.00	0.92	0.98	0.97	0.02	0.08	0.93	1.00
12	94.00	214.00	19.00	15.00	0.86	0.92	0.90	0.09	0.15	0.85	1.00
13	41.00	159.00	2.00	10.00	0.80	0.99	0.94	0.02	0.20	0.87	0.94
14	49.00	205.00	3.00	8.00	0.86	0.99	0.96	0.02	0.14	0.90	1.00
15	85.00	281.00	4.00	11.00	0.89	0.99	0.96	0.02	0.12	0.92	1.00
16	42.00	141.00	3.00	9.00	0.82	0.98	0.94	0.03	0.18	0.87	1.00
17	50.00	267.00	1.00	15.00	0.77	1.00	0.95	0.00	0.23	0.86	1.00
18	88.00	305.00	2.00	9.00	0.91	0.99	0.97	0.01	0.09	0.94	1.00

19	51.00	245.00	1.00	20.00	0.72	1.00	0.93	0.01	0.28	0.83	0.71
20	4.00	155.00	2.00	46.00	0.08	0.99	0.77	0.16	0.93	0.14	0.15
21	110.00	377.00	5.00	15.00	0.88	0.99	0.96	0.01	0.12	0.92	1.00
22	485.00	324.00	99.00	19.00	0.96	0.77	0.87	0.24	0.05	0.89	0.97
23	93.00	501.00	3.00	41.00	0.69	0.99	0.93	0.01	0.31	0.81	0.78
24	122.00	536.00	3.00	31.00	0.80	0.99	0.95	0.01	0.20	0.88	1.00
25	96.00	491.00	2.00	19.00	0.83	1.00	0.97	0.00	0.17	0.90	1.00
26	126.00	324.00	6.00	3.00	0.98	0.98	0.98	0.02	0.02	0.97	1.00
27	94.00	472.00	2.00	25.00	0.79	1.00	0.95	0.01	0.21	0.87	1.00
28	92.00	476.00	1.00	40.00	0.70	1.00	0.93	0.00	0.30	0.82	0.75
29	302.00	317.00	53.00	20.00	0.94	0.86	0.89	0.15	0.07	0.89	0.96
30	24.00	508.00	1.00	59.00	0.29	1.00	0.90	0.01	0.71	0.44	0.39
31	26.00	927.00	2.00	173.00	0.13	1.00	0.84	0.02	0.87	0.23	0.12
32	49.00	389.00	1.00	51.00	0.49	1.00	0.89	0.01	0.51	0.65	0.55
33	81.00	248.00	5.00	10.00	0.89	0.98	0.96	0.02	0.11	0.92	1.00
34	82.00	349.00	1.00	15.00	0.85	1.00	0.96	0.00	0.16	0.91	1.00
35	109.00	484.00	7.00	36.00	0.75	0.99	0.93	0.02	0.25	0.84	0.91
36	68.00	311.00	5.00	23.00	0.75	0.98	0.93	0.02	0.26	0.83	0.97
37	99.00	369.00	2.00	8.00	0.93	0.99	0.98	0.01	0.08	0.95	1.00
38	33.00	149.00	0.00	3.00	0.92	1.00	0.98	0.00	0.08	0.96	1.00
39	74.00	311.00	0.00	13.00	0.85	1.00	0.97	0.00	0.15	0.92	1.00
40	6.00	297.00	0.00	69.00	0.08	1.00	0.81	0.00	0.92	0.15	0.11
41	98.00	483.00	1.00	24.00	0.80	1.00	0.96	0.00	0.20	0.89	1.00
42	4.00	260.00	1.00	59.00	0.06	1.00	0.81	0.06	0.94	0.12	0.16
43	71.00	371.00	1.00	11.00	0.87	1.00	0.97	0.00	0.13	0.92	1.00
44	59.00	330.00	2.00	19.00	0.76	0.99	0.95	0.01	0.25	0.85	1.00
45	23.00	92.00	0.00	2.00	0.92	1.00	0.98	0.00	0.08	0.96	1.00
46	41.00	143.00	1.00	7.00	0.85	0.99	0.96	0.01	0.15	0.91	1.00
47	101.00	398.00	3.00	27.00	0.79	0.99	0.94	0.01	0.21	0.87	0.94
48	99.00	370.00	3.00	7.00	0.93	0.99	0.98	0.01	0.07	0.95	1.00
49	47.00	260.00	1.00	14.00	0.77	1.00	0.95	0.00	0.23	0.86	1.00
50	142.00	302.00	7.00	4.00	0.97	0.98	0.98	0.02	0.03	0.96	1.00

51	56.00	354.00	0.00	20.00	0.74	1.00	0.95	0.00	0.26	0.85	1.00
52	34.00	156.00	1.00	5.00	0.87	0.99	0.97	0.01	0.13	0.92	1.00
53	62.00	395.00	0.00	27.00	0.70	1.00	0.94	0.00	0.30	0.82	1.00
54	56.00	311.00	0.00	14.00	0.80	1.00	0.96	0.00	0.20	0.89	1.00
55	125.00	249.00	20.00	17.00	0.88	0.93	0.91	0.08	0.13	0.87	0.85
56	41.00	274.00	0.00	16.00	0.72	1.00	0.95	0.00	0.28	0.84	1.00
57	59.00	438.00	8.00	43.00	0.58	0.98	0.91	0.03	0.43	0.70	0.41
58	65.00	275.00	1.00	13.00	0.83	1.00	0.96	0.00	0.17	0.90	1.00
59	51.00	245.00	1.00	20.00	0.72	1.00	0.93	0.01	0.28	0.83	0.71
60	4.00	155.00	2.00	46.00	0.08	0.99	0.77	0.16	0.93	0.14	0.15

Table 13: Statistical Results for Speech Tracking Using Wavelet Based Dynamic Filter

Wavelet Based Adaptive Speech Tracking & Similarity Measurement Performance				
Evaluation Metrics		**Wavelet + Adaptive KF**	**Wavelet + Non-Adaptive SR**	**Wavelet + Adaptive KF (E & Spectral Centroid)**
Sensitivity		0.9907	0.9721	0.9951
Specificity		0.9841	0.9893	0.9654
Matching Accuracy		0.9869	0.9877	0.9764
1/LR+		0.0106	0.0108	0.0349
LR-		0.0107	0.0280	0.0052
F-Score		0.9833	0.9676	0.9683
Tracking Accuracy (%)		0.9983	98.8506	1
Type I Error	**μ**	0.0019	0.0010	0.0040
	σ	0.0070	0.0053	0.0104
Type II Error	**μ**	0.0086	0.0157	2.7232e-04
	σ	0.0460	0.0817	0.0011

Table 14: Statistical Results for Adaptive Speech Tracking Using Wavelet Based Dynamic Filter

Kalman Filter Based Adaptive Speech Tracking & Similarity Measurement performance (Wavelets)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	1.00	6.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

2	1.00	12.00	0.00	1.00	0.50	1.00	0.93	0.00	0.50	0.67	1.00
3	4.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
4	3.00	11.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
5	1.00	12.00	0.00	1.00	0.50	1.00	0.93	0.00	0.50	0.67	1.00
6	2.00	4.00	1.00	0.00	1.00	0.80	0.86	0.20	0.00	0.80	1.00
7	2.00	5.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
8	1.00	6.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
9	3.00	11.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
10	12.00	38.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
11	11.00	26.00	2.00	0.00	1.00	0.93	0.95	0.07	0.00	0.92	1.00
12	9.00	21.00	2.00	0.00	1.00	0.91	0.94	0.09	0.00	0.90	1.00
13	4.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
14	9.00	18.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
15	13.00	31.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
16	10.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
17	8.00	23.00	1.00	0.00	1.00	0.96	0.97	0.04	0.00	0.94	1.00
18	11.00	28.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
19	9.00	23.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
20	3.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
21	13.00	35.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.96	1.00
22	38.00	60.00	1.00	0.00	1.00	0.98	0.99	0.02	0.00	0.99	1.00
23	19.00	37.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
24	19.00	37.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
25	18.00	39.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
26	12.00	26.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
27	16.00	39.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
28	21.00	53.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
29	18.00	39.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
30	26.00	36.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

31	33.00	65.00	0.00	1.00	0.97	1.00	0.99	0.00	0.03	0.99	1.00
32	15.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
33	11.00	21.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
34	18.00	32.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
35	18.00	42.00	2.00	0.00	1.00	0.95	0.97	0.05	0.00	0.95	1.00
36	13.00	26.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
37	13.00	31.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
38	5.00	9.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
39	14.00	30.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
40	9.00	19.00	4.00	0.00	1.00	0.83	0.88	0.17	0.00	0.82	1.00
41	17.00	32.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.97	1.00
42	8.00	19.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
43	16.00	27.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
44	16.00	28.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
45	4.00	10.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
46	20.00	34.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
47	19.00	24.00	1.00	0.00	1.00	0.96	0.98	0.04	0.00	0.97	1.00
48	23.00	32.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.98	1.00
49	6.00	21.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
50	22.00	27.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
51	16.00	28.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
52	4.00	10.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
53	15.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
54	17.00	21.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
55	30.00	24.00	4.00	0.00	1.00	0.86	0.93	0.14	0.00	0.94	0.90
56	10.00	22.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
57	15.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
58	18.00	21.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
59	9.00	23.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

60	3.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

Table 15: Statistical Results for Static Framing Based Speech Tracking Using Dynamic Filter

Search Region Based Non-Adaptive Tracking & Similarity Measurement Performance (Wavelets)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	1.00	6.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
2	1.00	18.00	0.00	2.00	0.33	1.00	0.90	0.00	0.67	0.50	0.67
3	4.00	17.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
4	3.00	11.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
5	1.00	18.00	0.00	2.00	0.33	1.00	0.90	0.00	0.67	0.50	0.67
6	2.00	4.00	1.00	0.00	1.00	0.80	0.86	0.20	0.00	0.80	1.00
7	2.00	5.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
8	1.00	6.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
9	3.00	11.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
10	13.00	36.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
11	9.00	26.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
12	8.00	25.00	2.00	0.00	1.00	0.93	0.94	0.07	0.00	0.89	1.00
13	4.00	17.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
14	6.00	15.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
15	10.00	31.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.95	1.00
16	7.00	14.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
17	7.00	21.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
18	12.00	30.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
19	6.00	22.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
20	3.00	18.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
21	10.00	39.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
22	33.00	78.00	1.00	0.00	1.00	0.99	0.99	0.01	0.00	0.99	1.00
23	17.00	46.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
24	16.00	40.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

25	18.00	45.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
26	12.00	30.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
27	16.00	47.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
28	22.00	62.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
29	18.00	44.00	1.00	0.00	1.00	0.98	0.98	0.02	0.00	0.97	1.00
30	17.00	52.00	0.00	1.00	0.94	1.00	0.99	0.00	0.06	0.97	1.00
31	31.00	81.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
32	18.00	38.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
33	10.00	31.00	0.00	1.00	0.91	1.00	0.98	0.00	0.09	0.95	1.00
34	14.00	34.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.97	1.00
35	19.00	51.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
36	11.00	31.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
37	13.00	36.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
38	5.00	9.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
39	15.00	33.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.97	1.00
40	9.00	25.00	1.00	0.00	1.00	0.96	0.97	0.04	0.00	0.95	1.00
41	11.00	37.00	0.00	1.00	0.92	1.00	0.98	0.00	0.08	0.96	1.00
42	8.00	20.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
43	16.00	31.00	2.00	0.00	1.00	0.94	0.96	0.06	0.00	0.94	1.00
44	13.00	36.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
45	4.00	10.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
46	16.00	40.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
47	12.00	29.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.96	1.00
48	20.00	42.00	1.00	0.00	1.00	0.98	0.98	0.02	0.00	0.98	1.00
49	6.00	22.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
50	21.00	28.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
51	14.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
52	4.00	10.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
53	14.00	41.00	1.00	0.00	1.00	0.98	0.98	0.02	0.00	0.97	1.00

54	17.00	25.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
55	17.00	43.00	2.00	1.00	0.94	0.96	0.95	0.05	0.06	0.92	1.00
56	8.00	20.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
57	16.00	40.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
58	14.00	28.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
59	6.00	22.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
60	3.00	18.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

Table 16: Statistical Results Speech Tracking Using Dynamic Filter & Traditional Silence Removal

Adaptive Kalman Filter with Energy & Spectral Centroid Silence Removal Approach (Wavelets)											
Test Cases	**TP**	**TN**	**FP**	**FN**	**Sen**	**Spec**	**Acc**	**LR+**	**LR-**	**F-Score**	**Tracking Accuracy**
1	1.00	6.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
2	2.00	5.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
3	10.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
4	7.00	20.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
5	10.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
6	5.00	15.00	1.00	0.00	1.00	0.94	0.95	0.06	0.00	0.91	1.00
7	4.00	16.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
8	5.00	15.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
9	5.00	9.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
10	21.00	41.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
11	27.00	42.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
12	24.00	25.00	1.00	0.00	1.00	0.96	0.98	0.04	0.00	0.98	1.00
13	13.00	19.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
14	15.00	29.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
15	23.00	33.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
16	10.00	22.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
17	15.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
18	33.00	28.00	2.00	0.00	1.00	0.93	0.97	0.07	0.00	0.97	1.00

19	17.00	33.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
20	15.00	29.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
21	25.00	44.00	10.00	1.00	0.96	0.81	0.86	0.19	0.05	0.82	1.00
22	47.00	97.00	1.00	1.00	0.98	0.99	0.99	0.01	0.02	0.98	1.00
23	60.00	42.00	2.00	0.00	1.00	0.95	0.98	0.05	0.00	0.98	1.00
24	55.00	52.00	3.00	0.00	1.00	0.95	0.97	0.05	0.00	0.97	1.00
25	30.00	63.00	5.00	0.00	1.00	0.93	0.95	0.07	0.00	0.92	1.00
26	23.00	50.00	1.00	0.00	1.00	0.98	0.99	0.02	0.00	0.98	1.00
27	30.00	56.00	6.00	0.00	1.00	0.90	0.93	0.10	0.00	0.91	1.00
28	31.00	66.00	0.00	1.00	0.97	1.00	0.99	0.00	0.03	0.98	1.00
29	47.00	62.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
30	40.00	62.00	2.00	0.00	1.00	0.97	0.98	0.03	0.00	0.98	1.00
31	72.00	116.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
32	29.00	51.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
33	16.00	36.00	3.00	1.00	0.94	0.92	0.93	0.08	0.06	0.89	1.00
34	25.00	41.00	2.00	0.00	1.00	0.95	0.97	0.05	0.00	0.96	1.00
35	44.00	53.00	1.00	0.00	1.00	0.98	0.99	0.02	0.00	0.99	1.00
36	21.00	41.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
37	36.00	38.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
38	7.00	15.00	4.00	0.00	1.00	0.79	0.85	0.21	0.00	0.78	1.00
39	20.00	40.00	2.00	0.00	1.00	0.95	0.97	0.05	0.00	0.95	1.00
40	33.00	34.00	1.00	0.00	1.00	0.97	0.99	0.03	0.00	0.99	1.00
41	44.00	52.00	2.00	0.00	1.00	0.96	0.98	0.04	0.00	0.98	1.00
42	25.00	36.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.98	1.00
43	27.00	45.00	2.00	0.00	1.00	0.96	0.97	0.04	0.00	0.96	1.00
44	25.00	30.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.98	1.00
45	5.00	8.00	1.00	0.00	1.00	0.89	0.93	0.11	0.00	0.91	1.00
46	15.00	11.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
47	38.00	46.00	2.00	0.00	1.00	0.96	0.98	0.04	0.00	0.97	1.00

48	29.00	41.00	5.00	0.00	1.00	0.89	0.93	0.11	0.00	0.92	1.00
49	19.00	30.00	0.00	1.00	0.95	1.00	0.98	0.00	0.05	0.97	1.00
50	35.00	39.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
51	20.00	48.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
52	11.00	16.00	5.00	0.00	1.00	0.76	0.84	0.24	0.00	0.81	1.00
53	27.00	46.00	1.00	0.00	1.00	0.98	0.99	0.02	0.00	0.98	1.00
54	24.00	28.00	4.00	0.00	1.00	0.88	0.93	0.13	0.00	0.92	1.00
55	23.00	39.00	1.00	0.00	1.00	0.97	0.98	0.03	0.00	0.98	1.00
56	11.00	35.00	2.00	1.00	0.92	0.95	0.94	0.06	0.09	0.88	1.00
57	26.00	64.00	2.00	0.00	1.00	0.97	0.98	0.03	0.00	0.96	1.00
58	21.00	35.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
59	17.00	33.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00
60	15.00	29.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00

Appendix C: Simulation Tools

1. PRAAT

PRAAT (Resource: http://www.fon.hum.uva.nl/paul/praat.html) is a special tool that has been used for speech annotations, speech analysis, speech synthesis, speech manipulation, labelling and segmentation, and for learning the algorithms. PRAAT has the ability to analyse different speech features including formant frequency analysis, pitch tracking, speech intensity and spectrogram. Moreover, it provides a user friendly interface for graphical representation of these features. Figures below demonstrate the use of PRAAT for analysing speech signal for different purpose.

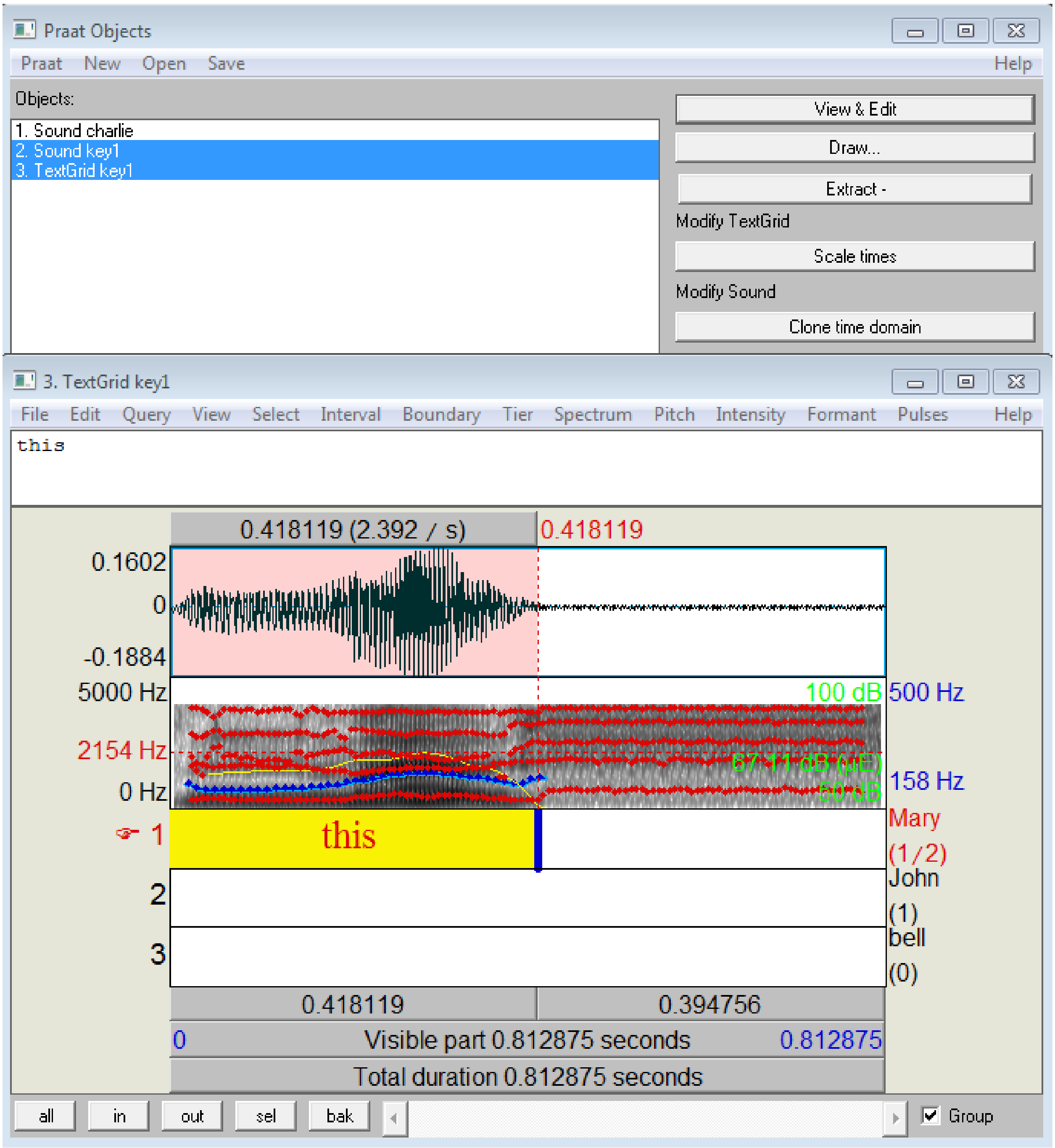


Figure 1: Pitch, Formants, Frequency Spectrum, and Energy Analysis using PRAAT

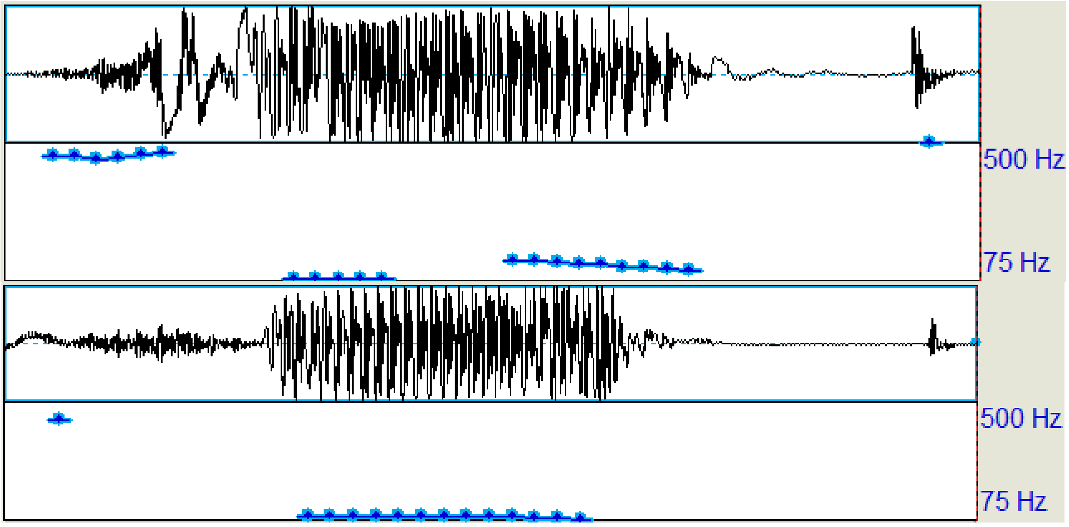


Figure 2: Pitch Track Analysis Using PRAAT For Speech Utterance ‘Short’

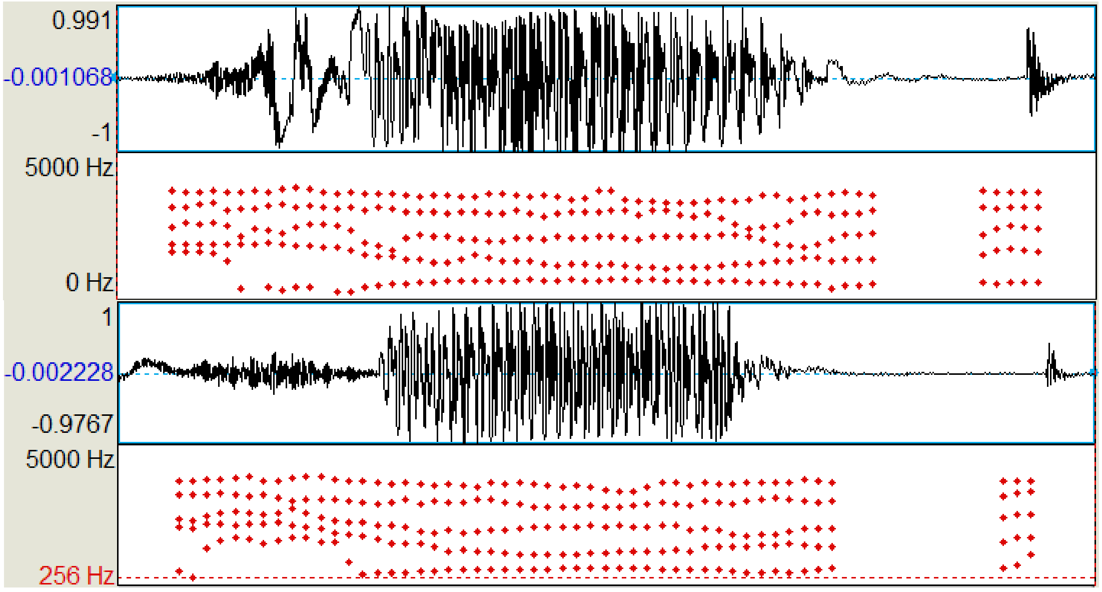


Figure 3: Formants Analysis Using PRAAT For Speech Utterance ‘Short’

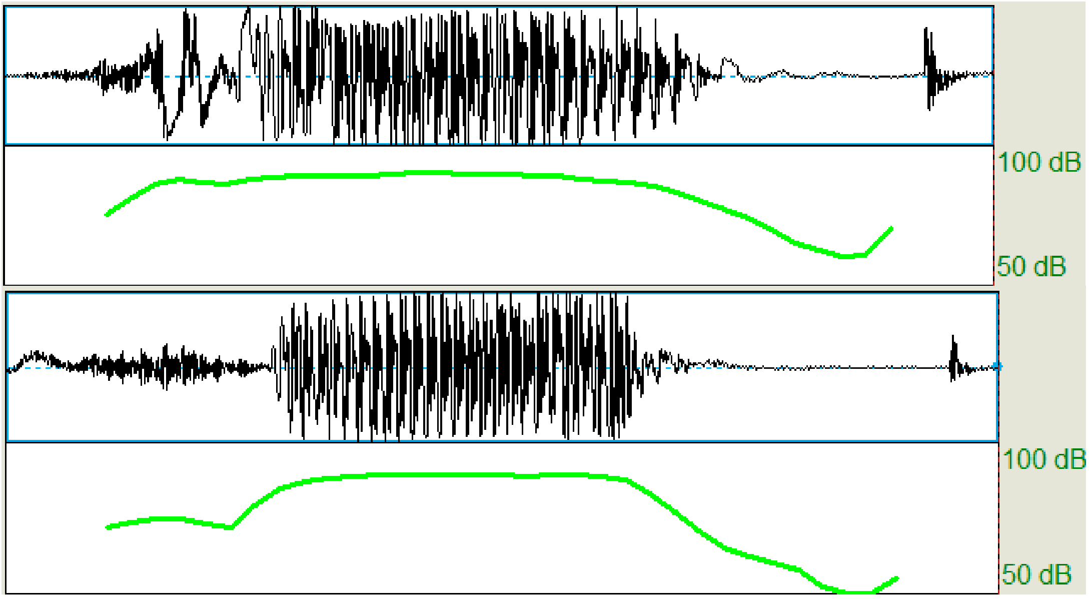
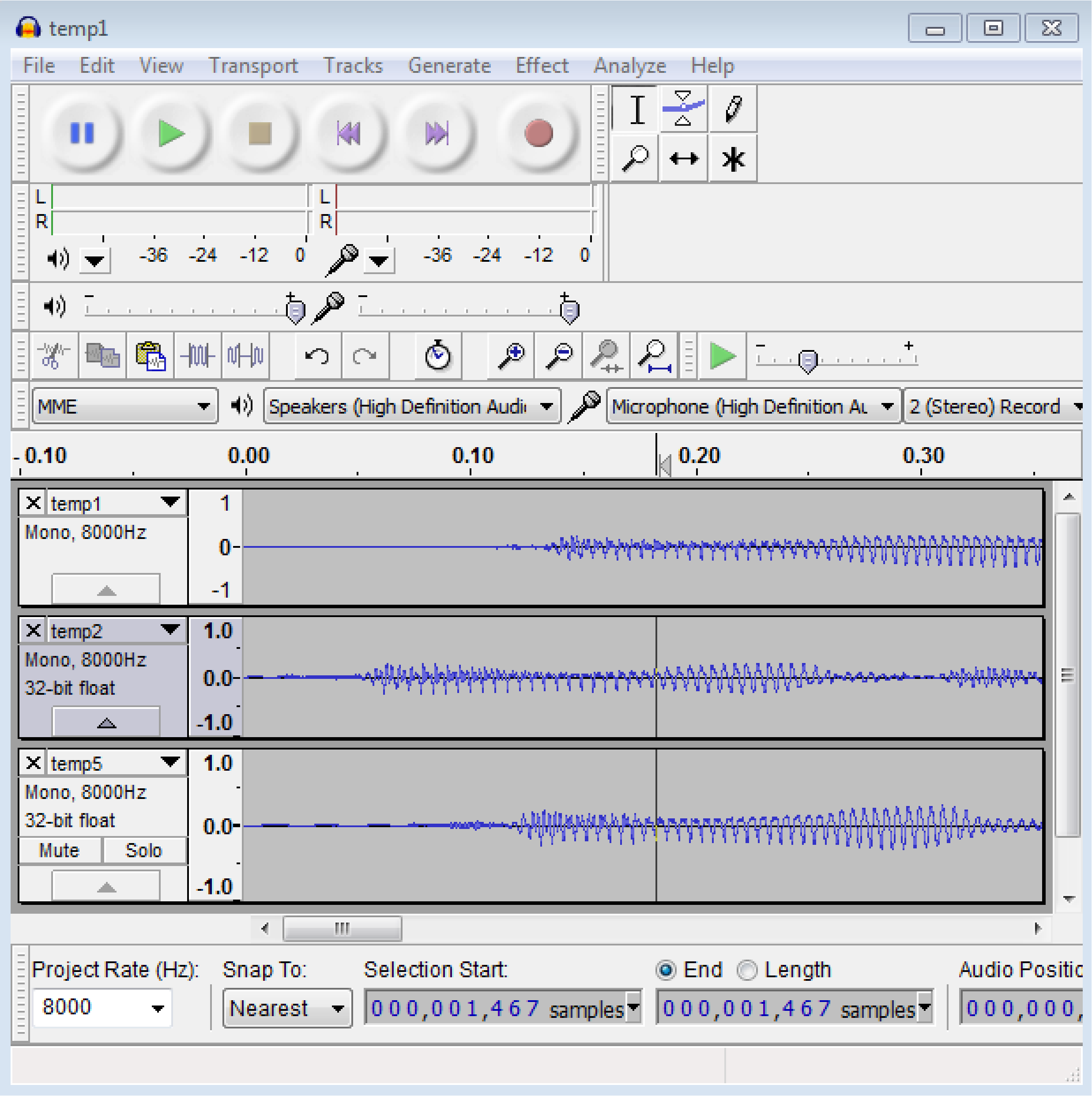


Figure 4: Intensity Distrubution Analysis Using PRAAT For Speech Utterance ‘Short’

2. Audacity

Audacity (Resource: http://sourceforge.net/projects/audacity/) is a free, open source, cross-platform that is normally used for recording and editing sounds. It may be utilized for speech editing in terms of sample rate conversion, noise reduction, silence removal, speech effects, tempo editing, channel manipulation and much more. Figure below provides a sketch of the user interface for Audacity.



MME

Speakers (High Definition Audi

Microphone (High Definition Au

2 (Stereo) Record

temp1
Mono, 8000Hz

temp2
Mono, 8000Hz
32-bit float

temp5
Mono, 8000Hz
32-bit float

Project Rate (Hz): 8000

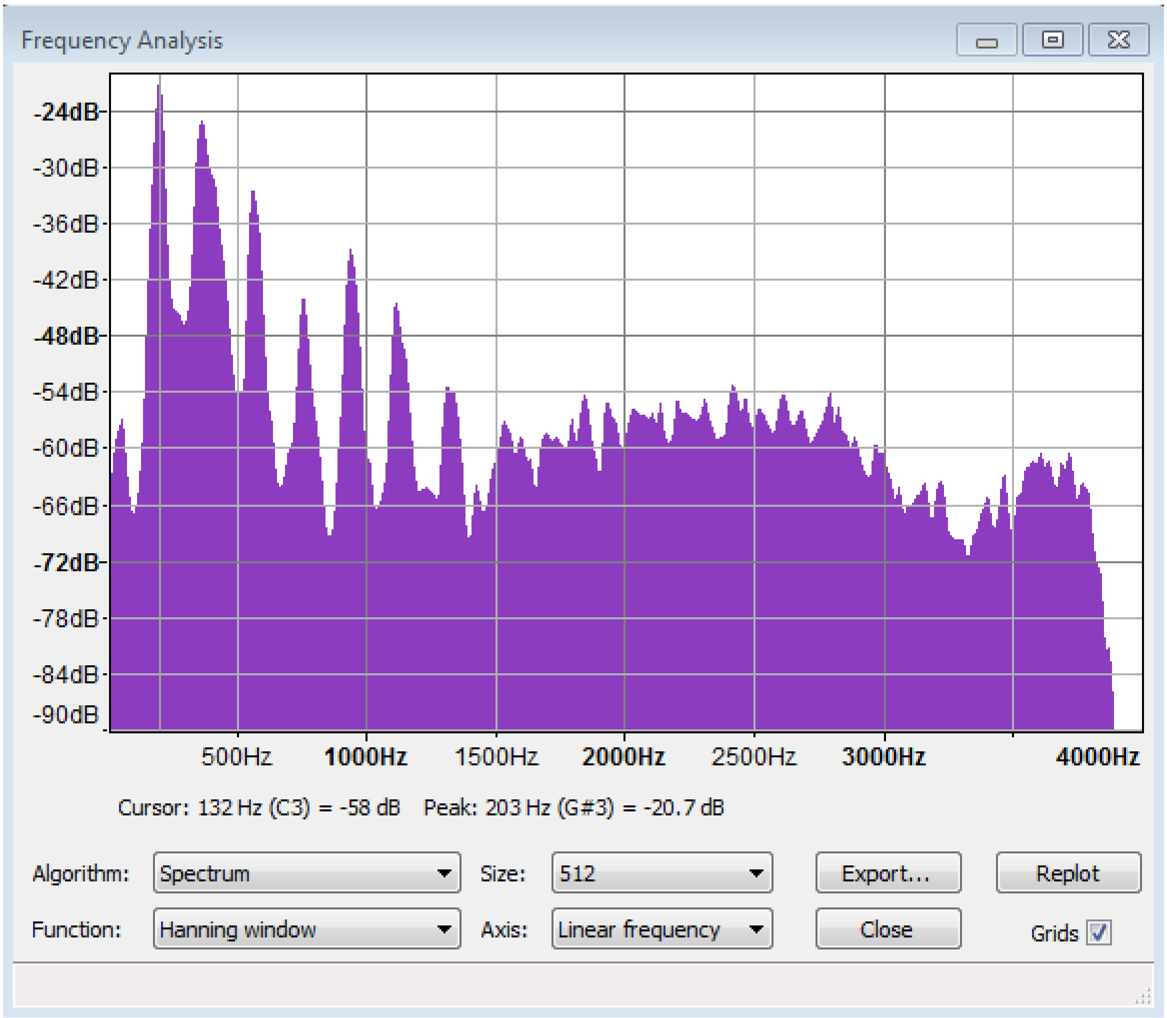
Snap To: Nearest

Selection Start: 000,001,467 samples

● End ○ Length
000,001,467 samples

Audio Positic
000,000,

Figure 5: Exporting Multiple Outputs From Audacity Platform



Cursor: 132 Hz (C3) = -58 dB   Peak: 203 Hz (G#3) = -20.7 dB

Algorithm: Spectrum   Size: 512

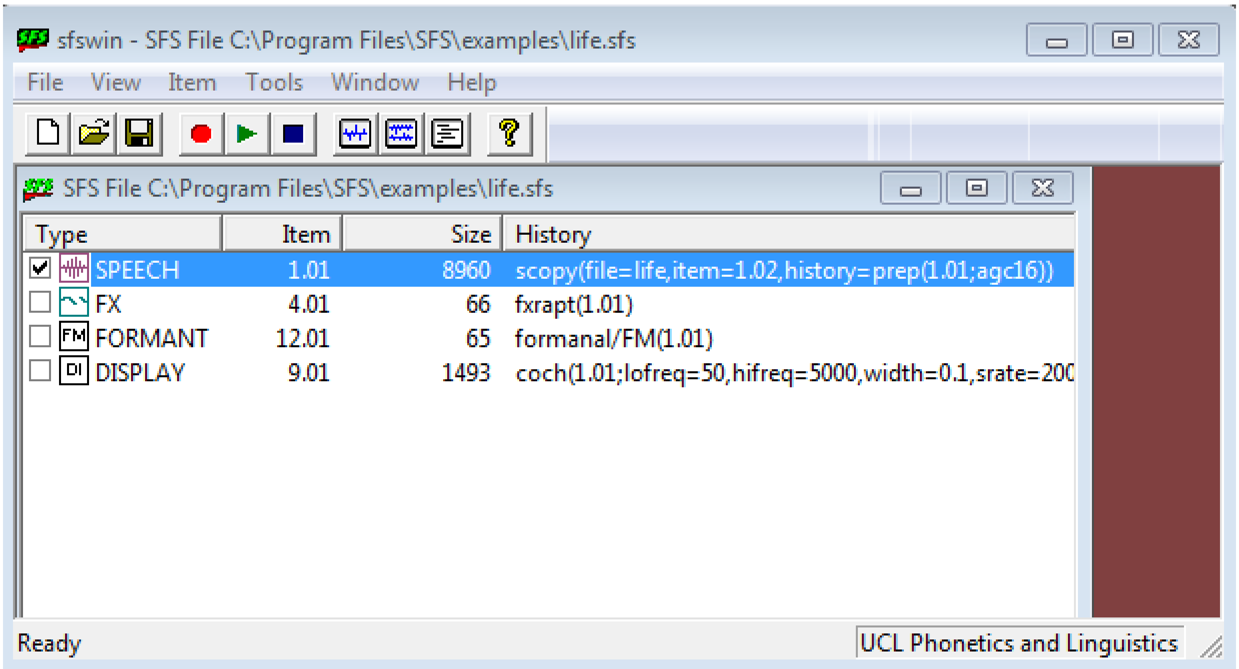
Function: Hanning window   Axis: Linear frequency

Grids ☑

Figure 6: Analysing Frequency Spectrum of a Signal using Audacity Platform

3. Sppech Filing System (SFS)

Speech filing system (*https://www.phon.ucl.ac.uk/resource/sfs/download.php*) has been used as a research tool for speech processing that is developed by University College. The SFS provides a large functionality set in the form of executable files that can be export easily for the reusability in cross platforms. It is an open resource platform to conduct the speech processing related research that include software tools, file and data formats, subroutine libraries, graphics, special programming languages and tutorial documentation. Operations like acquisition, replay, display and labelling, spectrographic and formant analysis and fundamental frequency estimation are the most common examples of the SFS functionalities. Except analysing of speech signal, it provides a large body of libraries for signal processing, synthesis and recognition that can be reused in cross platforms for personal software development.

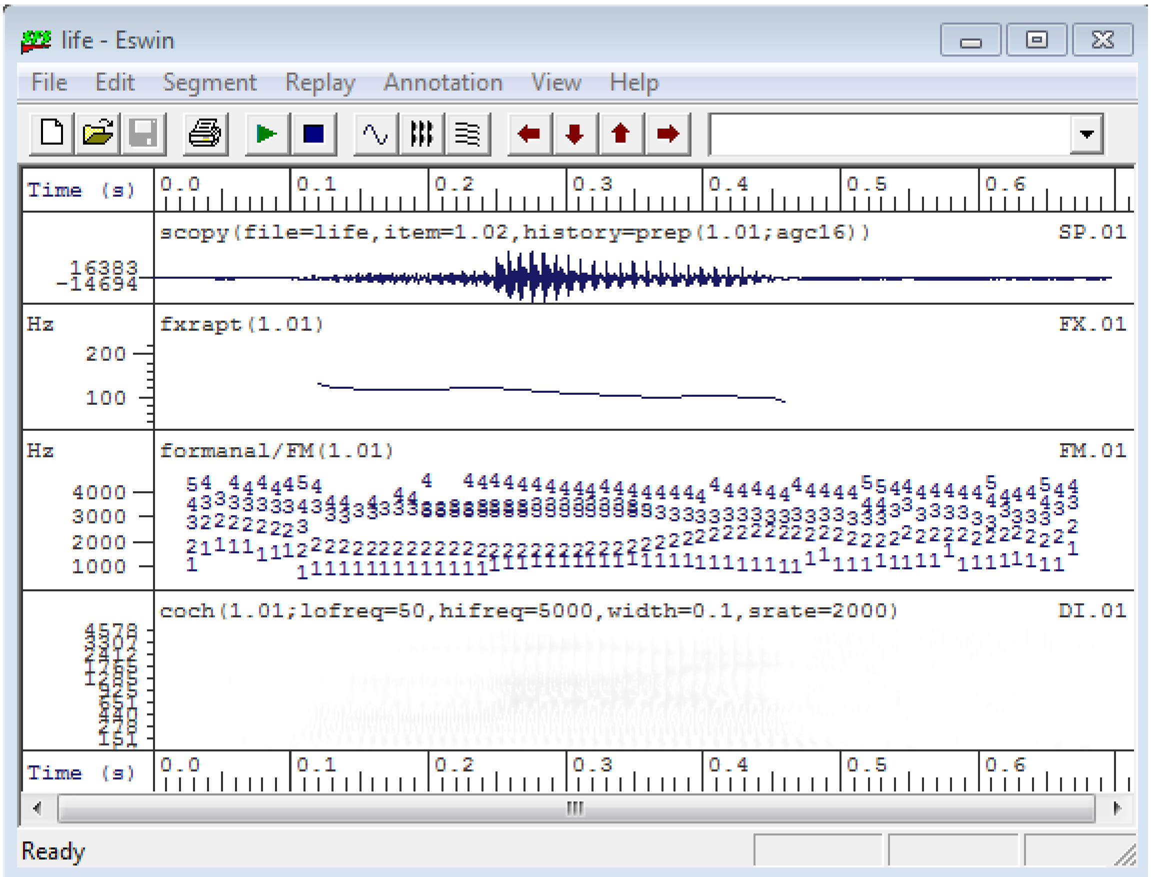

sfswin - SFS File C:\Program Files\SFS\examples\life.sfs
File View Item Tools Window Help
SFS File C:\Program Files\SFS\examples\life.sfs

Type	Item	Size	History
☑ SPEECH	1.01	8960	scopy(file=life,item=1.02,history=prep(1.01;agc16))
□ FX	4.01	66	fxrapt(1.01)
□ FORMANT	12.01	65	formanal/FM(1.01)
□ DISPLAY	9.01	1493	coch(1.01;lofreq=50,hifreq=5000,width=0.1,srate=200

Ready

UCL Phonetics and Linguistics

Figure 7: SFS GUI for Speech Analysis



```
scopy(file=life,item=1.02,history=prep(1.01;agc16))    SP.01
fxrapt(1.01)                                            FX.01
formanal/FM(1.01)                                       FM.01
coch(1.01;lofreq=50,hifreq=5000,width=0.1,srate=2000)   DI.01
Ready
```

Figure 8: Formants Analysis using SFS Platform

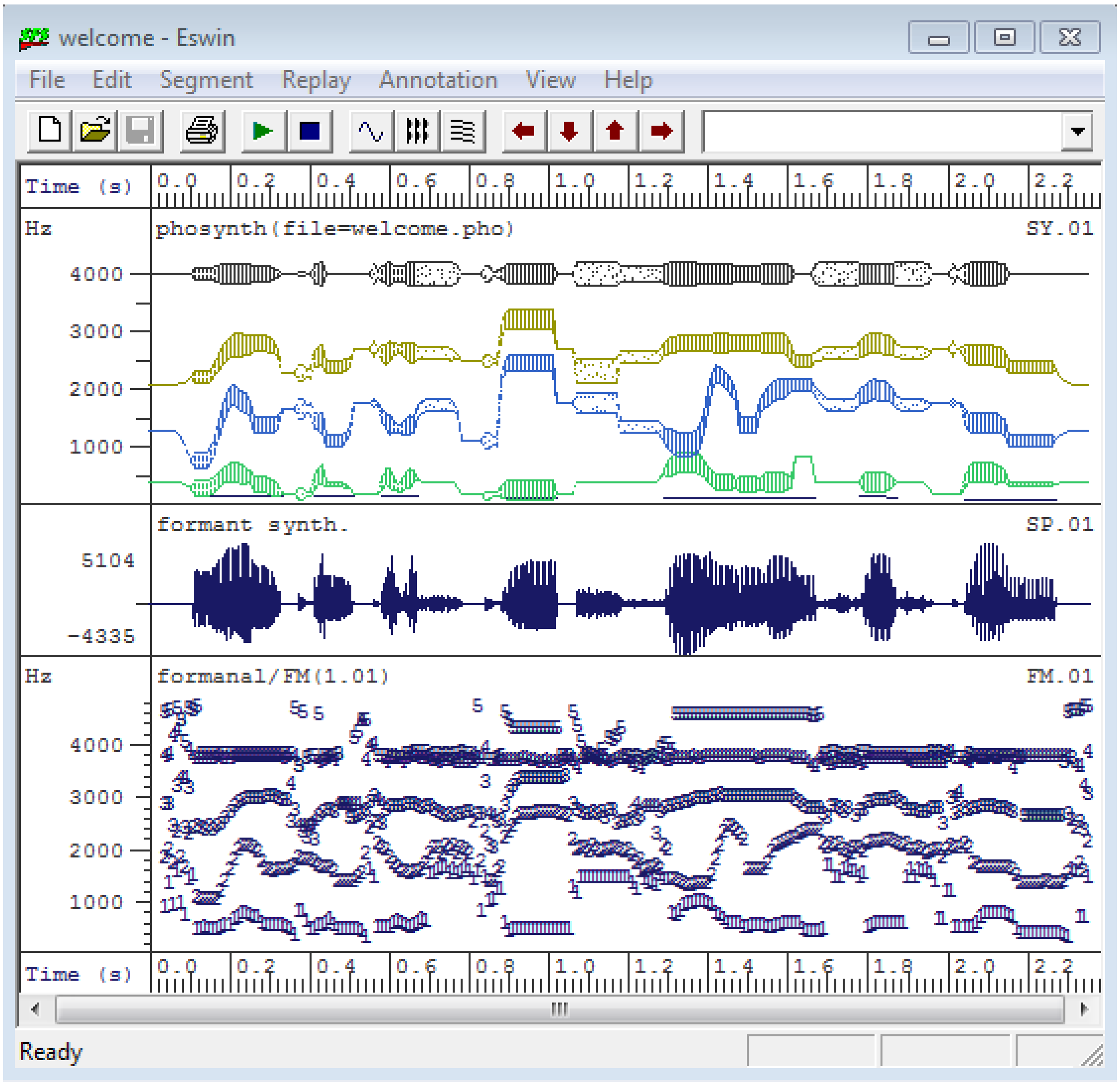


Figure 9: Formants Analysis using SFS Platform

Appendix D: Matlab Scripts

CD consisting Matlab scripts for following functionalities:

- ✓ Time Warped Continuous Speech Tracking (TWCST)
- ✓ Key-word Spotting
- ✓ Silence Removal
- ✓ Feature Extraction
- ✓ Dynamic Time Warping
- ✓ Dynamic State Model and Kalman Filter