ARAB ACADEMY FOR SCIENCE, TECHNOLOGY AND MARITIME TRANSPORT

College of Engineering and Technology

Department of Electronics and Communications Engineering

Robust Automatic Speech Recognition Systems Based on Using Adaptive Time-Frequency Techniques

By

Ahmed Mostafa Ahmed Abdelmaguid Gouda

A thesis submitted to AASTMT in partial

Fulfillment of the requirements for the award of the degree of

MASTER'S OF SCIENCE

in

ELECTRONICS AND COMMUNICATIONS ENGINEERING

Supervisors

Prof. Mohamed Essam Khedr
Electronics and Communications Engineering Department
AASTMT – Alexandria

Dr. Mohamed Essam Tamazin
Electronics and Communications Engineering Department
AASTMT – Alexandria

2017

Acknowledgments

Firstly, Thanks to Almighty Allah for guiding me and giving me the ability to complete my studies.

I am deeply indebted to Doctor Mohamed Tamazin and Professor Mohamed Khedr for their creative and eclectic advices. I appreciate their help by which my studies could be complete successfully.

It has been honorable to be taught by Professor Mohamed Khedr, who has enriched my knowledge since my bachelor and master studies. He has established the bases which I needed for excelling in my degrees. It is also important to mention that he has instilled my eagerness to learn by his availability for questions and bringing up thought-provoking topics for discussion.

I would also like to express my gratitude to Doctor Mohamed Tamazin for his efforts and whose help in every aspect of my research has been valuable for the completion of my studies. He has developed my skills in researching. His precision and his constant concern for the accuracy of this research have added to its quality.

Last but not least, I dedicate all my studies and work to my late father whom I wish him to be proud of me. Without his persistence and insistence, I would not have taken this path. I owe my wholehearted thanks and appreciation to Mother for having supported me through this journey and for her extreme patience, love, and encouragement. Thanks to my sisters, for always being there for me and motivating me to take new steps every day.

Abstract

Many of the new waves of consumer-centric applications are based on using Automatic Speech Recognition (ASR) systems, such as voice command interfaces, speech-to-text applications, and data entry processes. Although the ASR systems have extremely improved in recent decades, the speech recognition system performance still significantly degrades in the presence of noisy environments. Developing robust ASR system can work in real-world noise and other acoustic distorting conditions is an attractive research topic. Many algorithms have been developed to deal with this problem, most of these algorithms are based on modeling the behavior of the human auditory system with perceived noisy speech. In this research, two proposed systems are implemented. In the first proposed system, the Mel-Frequency Cepstral Coefficients (MFCC) is modified to robust against the noise, where the spectrogram is used as a time-frequency analysis tool. The proposed system is designed to decrease the energy values which is highly affected by the noise. It uses an adaptive filtering technique to robust against the noise without loss of performance in case of undistorted speech data. The proposed system is evaluated in the presence of Additive White Gaussian Noise (AWGN). The experimental results have demonstrated that the proposed MFCC method provides significant improvements in recognition accuracy at different Signal to Noise Ratios (SNR). For instance, the recognition accuracy has improved by 34.45% and 20.37% in comparison to standard MFCC and RASTA-PLP at SNR 0dB, respectively. In the second proposed system, the Power-Normalized Cepstral Coefficients (PNCC) system is modified to robust against the different types of noise, where a new technique based on Gammatone channels filtering combined with a channel bias minimizing is used to suppress the noise. The performance of the proposed system is evaluated in the presence of Additive White Gaussian Noise (AWGN) and seven different types of environmental noises. The experimental results have shown that the proposed method provides significant improvements in recognition accuracy significant improvements at low Signal to Noise Ratios (SNR). The highest improvement in recognition rate in comparison to MFCC and RASTA-PLP methods are obtained in Subway noise condition at SNR 5dB. It is improved by 55.72% and 47.87% more than MFCC and RASTA-PLP methods, respectively. However, the highest enhancement in recognition rate in comparison to GFCC and PNCC methods are obtained in the case of Car noise. It is enhanced by 40.02% in comparison to GFCC method at SNR 0dB, while it is improved by 19.51% in comparison to PNCC method at SNR -5dB.

Table of Contents

Contents Pages

Acknowledgments i

Abstract ii

Table of Contents iii

List of Tables v

List of Figures vi

List of Abbreviations ix

List of Symbols xi

1 Chapter ONE: Introduction 2

1.1 Introduction 2

1.2 Problem Definition 3

1.3 Research Objectives 3

1.4 Thesis Outline 3

2 Chapter TWO: Background and Literature Review 6

2.1 Automatic Speech Recognition Approaches 6

2.1.1 Feature-space Approaches 6

2.1.2 Model-space Approaches 7

2.2 Automatic Speech Recognition System Architecture 8

2.2.1 Training Stage 8

2.2.2 Testing Stage 14

2.3 Feature Extraction Techniques 17

2.3.1 Mel-Frequency Cepstral Coefficients (MFCC) 17

2.3.2 RelAtive SpecTrAl Perceptual Linear Predictive (RASTA-PLP) 21

2.3.3 Gammatone Frequency Cepstral Coefficients (GFCC) 24

2.3.4 Power-Normalized Cepstral Coefficients (PNCC) 27

Table of Contents (Cont'd)

2.3.5 Cepstral Mean Normalization (CMN) 33

2.3.6 Energy Features 34

2.3.7 Dynamic Features Delta (Δ) and Delta-delta (ΔΔ) 35

2.4 Performance Evaluation 35

3 Chapter THREE: Proposed Method 37

3.1 Proposed Method 1: Modified MFCC Technique 37

3.1.1 SNR Estimation 38

3.1.2 Time-frequency Mask Mapping 39

3.1.3 Mean Smoothing Filter 41

3.2 Proposed Method 2: Modified PNCC Technique 43

3.2.1 Medium Time Average Filtering 44

3.2.2 Channel Bias Minimizing 45

4 Chapter FOUR: Experimental Works and Results 49

4.1 Database Description 49

4.2 Experimental Results of Proposed Method 1 49

4.3 Experimental Results of Proposed Method 2 51

5 Chapter Five: Conclusion and Future Work 70

5.1 Conclusion 70

5.2 Future Work 72

References 73

List of Tables

Table 4-1: Percentage Word Recognition Rate (WRR) for AWGN in the first proposed method 51

Table 4-2: Noise description 52

Table 4-3: Percentage Word Recognition Rate (WRR) for AWGN in the second proposed method 55

Table 4-4: Percentage Word Recognition Rate (WRR) for Airport noise 56

Table 4-5: Percentage Word Recognition Rate (WRR) for Babble noise 58

Table 4-6: Percentage Word Recognition Rate (WRR) for Car noise 59

Table 4-7: Percentage Word Recognition Rate (WRR) for Exhibition noise 61

Table 4-8: Percentage Word Recognition Rate (WRR) for Restaurant noise 62

Table 4-9: Percentage Word Recognition Rate (WRR) for Street noise 64

Table 4-10: Percentage Word Recognition Rate (WRR) for Subway noise 65

List of Figures

Figure 2-1: Block diagram of training stage for ASR system 9

Figure 2-2: Markov states modeling for word "FIVE" 10

Figure 2-3: Baum-welch Forward-Backward algorithm 13

Figure 2-4: Block diagram of testing stage for ASR system 15

Figure 2-5: Viteribi decoder 16

Figure 2-6: Block diagram of MFCC system 17

Figure 2-7: Magnitude response of pre-emphasize filter 18

Figure 2-8: Windowing speech wave form 18

Figure 2-9: Frequency domain Mel-frequency scale relation 20

Figure 2-10: Triangular Mel-filter banks 20

Figure 2-11: Block digram of RASTA-PLP system 21

Figure 2-12: Bark filter banks 22

Figure 2-13: Block diagram of GFCC system 24

Figure 2-14: Physiology of the cochlea in the human auditory system 25

Figure 2-15: Set of 25 gammatone impulse response with center frequencies from 100 Hz to 4 KHz 26

Figure 2-16: Gammatone filter banks 26

Figure 2-17: Block diagram of PNCC system 27

Figure 2-18: Normalized 40 Gammatone filter banks 28

Figure 2-19: Block diagram of Asymmetric Noise Suppression with Temporal Masking . 29

Figure 2-20: Block diagram of Temporal Masking 31

Figure 2-21: Moving cepstral features to have a zero mean 34

Figure 3-1: Block diagram of the proposed MFCC method 37

Figure 3-2: Adaptive time-frequency masking block diagram 37

Figure 3-3: The threshold values at different Estimated Signal to Noise Ratios (ESNR) .. 40

List of Figures (Cont'd)

Figure 3-4: Spectrogram of the uttered word 'one' at SNR = 5 dB 42

Figure 3-5: Block diagram of the second proposed system 43

Figure 3-6: Normalized 25 Gammatone filter banks..... 44

Figure 3-7: Recognition performance at different filter average width at 5dB 46

Figure 3-8: Recognition performance at different filter average width at -5dB..... 46

Figure 4-1: Feature extraction stage for the first experiment 50

Figure 4-2: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for AWGN in the first proposed method 50

Figure 4-3: Feature extraction stage of the second experiment..... 53

Figure 4-4: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for AWGN in the second proposed method..... 54

Figure 4-5: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Airport noise..... 56

Figure 4-6: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Babble noise 57

Figure 4-7: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Car noise..... 59

Figure 4-8: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Exhibition noise..... 60

Figure 4-9: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Restaurant noise..... 62

Figure 4-10: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Street noise 63

Figure 4-11: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Subway noise..... 65

Figure 4-12: Percentage improvement rate for all types of noise at SNR -5dB..... 66

Figure 4-13: Percentage improvement rate for all types of noise at SNR 0dB 67

List of Figures (Cont'd)

Figure 4-14: Percentage improvement rate for all types of noise at SNR 5dB 68

List of Abbreviations

ALSD	Average Localized Synchrony Detection
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
AWGN	Additive White Gaussian Noise
BN	Bottle-Neck
CD	Context-Dependent
CMN	Cepstral Mean Normalization
CMU	Carnegie Mellon University
CMVN	Cepstral Mean and Variance Normalization
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
FFT	Fast Fourier Transform
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
GSNR	Global Signal to Noise Ratio
HEQ	Histogram EQualization
HMM	Hidden Markov Model
IFFT	Inverse Fast Fourier Transform
LVSR	Large-Vocabulary Speech Recognition
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
PLP	Perceptual Linear Predictive
PMVDR	Perceptual Minimum Variance Distortionless Response
PNCC	Power-Normalized Cepstral Coefficients

List of Abbreviations (Cont'd)

PSD	Power Spectral Density
RASTA	RelAtive SpecTrAl
SNR	Signal to Noise Ratio
SPARK	Sparse Auditory Reproducing Kernel
TI	Texas Instrument
VAD	Voice Activity Detection
WRR	Word Recognition Rate
ZCPA	Zero Crossing Peak Amplitude

List of Symbols

T	Number of frames in a speech sequence
O	Sequence of feature vector or observations ($O = o_1, o_2, o_3, ..., o_T$)
N	Number of Markov states
Q	Set of Hidden Markov states ($Q = q_1, q_2, q_3, ..., q_N$)
$\widehat{M}$	Number of the mixture parameters
$\tilde{c}_{j\hat{m}}$	Weight of $\hat{m}^{th}$ parameter
B_i	Observation likelihoods or emission probabilities ($b_i(o_t)$)
$N(.;\mu,\Sigma)$	Multivariate Gaussian with mean vector μ and covariance matrix Σ
$L_j(t)$	Probability of being in state j at time t.
$\alpha_j(t)$	Forward probability
$\beta_j(t)$	Backward probability
a_{ij}	Transition probability
P	Total probability
$v_t(j)$	Viterbi trellis element
n	Sample index
$s[n]$	Speech samples in time domain
α	Pre-emphasizing filter coefficient
$y[n]$	Pre-emphasized samples
J	Total number of samples within frame
$w[n]$	Hamming window samples
k	Spectrum index
$Y(k)$	Frequency domain of values

List of Symbols (Cont'd)

f	Frequency scale
f_{mel}	Mel-scale
l	Filter bank number or channel
$\Gamma_l(k)$	Triangular Mel-filter bank function
K	Spectrum resolution
$E(l)$	Mel-spectrum value
c_i'	Cepstral feature value
L	Total numbers of features
ω	Angular frequency in rad/s
Ω	Bark frequency scale
$\Psi(\Omega)$	Critical band masking curve for Bark filter
$\Theta(\Omega_i)$	Bark spectrum value
$H(z)$	Filter transfare function
$E_{pre}(\omega)$	Pre-emphasize equal-loudness
$\Phi(\Omega)$	Intensity loudness power low
f_c	Gammatone center frequency,
ϕ	Phase of the Gammatone carrier,
$\hat{a}$	Amplitude of Gammatone
g	Gammatone filter order
$\hat{b}$	Gammatone filter bandwidth
m	Frame index
$G_l(k)$	Gammatone filter function

List of Symbols (Cont'd)

$P[m,l]$	Short-time spectral power
M	Half average filter width for Medium-time
$\tilde{Q}[m,l]$	Medium-time power
λ_a, λ_b	Constants between zero and one
λ_t	Forgetting factor equal to 0.85
$\tilde{R}[m,l]$	Asymmetric Noise Suppression with Temporal Masking values
c	Fixed threshold equal to 2
$\tilde{S}[m,l]$	Spectral Weight Smoothing values
$\tilde{T}[m,l]$	Time-Frequency normalization values
$\mu[m]$	Mean power estimated values
λ_μ	Forgetting factor equal to 0.999
k_{const}	Arbitrary constant equal to 4×10^7
$U[m,l]$	Mean power normalization values
$V[m,l]$	Power function nonlinearity function
$\tilde{m}_i$	Mean value for features along frames
$c_n'(m)$	Cepstral mean normalized values
$\Delta d(t)$	Velocity variation features along frames
$\Delta\Delta d(t)$	Acceleration variation features along frames
N_w	Total number of substitution error
S	Number of deletion error
D	Number of deletion error
I	Number of insertion error

List of Symbols (Cont'd)

Acc	Percentage Word Accuracy
WRR	Percentage Word Recognition Rate
σ_s^2	Power of the signal
σ_n^2	Power of the noise
$vad[n]$	Detected voice activity within the speech waveform
$n[n]$	Noise samples in time domain
$I'(m,k)$	Two dimension mean filter kernel function
$I'_{norm}(m,k)$	Normalized two dimension mean filter kernel function
a_{fit}, b_{fit}	Constant values of the fit curve equation
θ	Channel Bias Minimization factor between 0 and 1
$\breve{Q}[m,l]$	Channel Bias Minimized values

CHAPTER 1

INTRODUCTION

1 Chapter ONE: Introduction

1.1 Introduction

Speech is one of the most important way to exchange information among people. Over decades, speech processing has been an attractive research topic, which motivated many researchers to build interface systems that can analyze and execute speech commands. Creating an Automatic Speech Recognition (ASR) system that can convert the acoustic signal to a string of words was a very challenging task in the past.

ASR systems have witnessed a remarkable development on a great diversity of tasks with increasing the usage scales and complexity. During the 1970s, the probabilistic paradigm Hidden Markov Model (HMM) [1] is considered a significant theoretical breakthrough in the development of speech recognition systems. The development of the HMMs theory besides fast development of computer hardware and algorithms contributed to making the continuous speech Recognition systems become the main research interest. By the end of the 1980s to 1990s, U.S. governmental institution whose name is Defense Advanced Research Project Agency (DARPA) has marked a few important milestones in ASR technology [2]. Thus, the speech recognition vocabulary has increased from a few hundred words to thousands of words. This evaluation has demonstrated a significant change in pushing ahead the ASR research field.

Despite increasing the recognized vocabulary size, there are other problems which appeared in other aspects during implementing the ASR practically and in realistic applications. Environmental noise is one of the main problems that affects the ASR systems by degrading its performance at different levels of noise. Besides the environmental noise, the performance of ASR systems is also affected by many problems such as channel bandwidth, channel reverberation, speaker dialect adaptation and sentence modeling. Building a robust speech recognition system that can provide services for millions of users is still a challenging research area.

Recently, the speech recognition market size has grown rapidly because of increasing research adoption by research organizations and huge companies. ASR systems are applied in diverse applications that include voice command interfaces, speech-to-text applications, and data entry process. These applications serve many sectors such as military, robotics, automotive, home automation, emergency applications, telephony services, security systems, healthcare systems, disabled services, multimedia systems and mobile applications.

1.2 Problem Definition

The performance of an Automatic Speech Recognition system is degraded in the presence of environmental noise. Designing a robust speech recognition system that can work at different several noise levels of non-stationary noisy environments faces several challenges. For instance, the system should be designed to robust against the noise at low SNR without degrading the performance in case of undistorted speech data. This can be achieved by suppressing the noise without removing the speech data information.

1.3 Research Objectives

The main objective of this thesis is to develop feature extraction techniques to improve ASR system robustness at various noisy environments. This objective can be achieved by:

- Developing a proposed method to suppress different environmental noises at different Signal to Noise Ratios (SNRs).
- Evaluating the performance of the proposed methods with the state-of-the-art techniques in the terms of Word Recognition Rate.

1.4 Thesis Outline

This thesis is organized as follows, Chapter 2 is a literature review on automatic speech recognition techniques which is divided in four sections. The first section in this chapter explains the approaches of the noise-robust Automatic Speech Recognition systems. This section is divided into two sub-sections feature-based approaches and model-based approaches. In the second section, the Automatic Speech Recognition system architecture for feature-based approaches systems is illustrated. The next section is an over view feature extraction techniques that are used in developing different systems in this research. In the last section, the methods that are used in evaluating the system performance are explicated.

Inspired by previous work, two proposed methods based on feature extraction approach were developed and explicated in Chapter 3. Both of the proposed methods were illustrated graphically and numerically.

Finally, Chapter 4 is experimental works that are implemented and the results that obtained to validate both of the proposed techniques and it is divided into three sections. The first section is a description of the TIDIGITS database which is used in this research. The second section is the

experimental works and the obtained results of the first proposed method, Similarly, The third section is the experimental works and the obtained results of the second proposed method.

CHAPTER 2

BACKGROUND AND LITREATURE REVIEW

2 Chapter TWO: Background and Literature Review

2.1 Automatic Speech Recognition Approaches

As stated in reference[3], the noise-robust Automatic Speech Recognition (ASR) Systems are classified into two main approaches. These approaches are Feature-Space approaches and Model-Space approaches. Each approach is based on building the particular system structure to suppress the noises which are explained in details in the following:

2.1.1 Feature-space Approaches

In the feature approaches, Hidden Markov Model (HMM) is usually used as a machine learning tool. These features mainly depend on robust against the noise to adapt the extracted feature in order to match the training features. These approaches are divided into three subcategories:

2.1.1.1 Noise Resistant features

Noise-resistant feature methods focus on decreasing the sensitivity of speech in the presence of the environmental noise. It depends on decrease the noise effect rather than remove the noise. The advantage of this technique no noise statistics estimations required.

a) Auditory based features

This mainly focuses on modeling the behavior of human auditory system toward the environmental noise. Mel-Frequency Cepstral Coefficients (MFCC) is one of the most widely used techniques[3] that models the human auditory system to a set of overlapped triangular filters banks warped to Mel-scale. The most researchers in the litreture evaluate their systems in comparison with MFCC. Perceptual Linear Predictive (PLP) [4] is also considered one of the important techniques. RelAtive SpecTrAl (RASTA) [5] filtering is combined with PLP system to removes channel noise that varies slowly compared to the speech waveform.

There are a many other feature extraction methods, which are used in the litreature such as Gammatone Frequency Cepstral Coefficients (GFCC) [6], Zero Crossing Peak Amplitude (ZCPA) [7], Average Localized Synchrony Detection (ALSD) [8], Perceptual Minimum Variance Distortionless Response (PMVDR) [9], Sparse Auditory Reproducing Kernel (SPARK) [10], Gabor filter bank features [11] and Power-Normalized Cepstral Coefficients (PNCC) [12]. All of this system were designed based on utilizing some auditory behaviors in analyzing the distorted speech waves. But there are not known standard defines which system is better than the other.

b) Neural Network Approaches

Artificial Neural Network (ANN) is a well-known approach that provides effective features for speech recognition systems, such as ANN-HMM [13] hybrid systems. The ANN systems usually designed based on Multi-Layer Perceptron (MLP) structure with a single hidden layer. The TANDEM [14] system was designed to combine ANN system with (Gaussian Mixture Model) GMM and it demonstrated a high performance on noisy datasets.

Bottle-Neck (BN) [15] features were designed based on ANN feature extraction method using a five-layer MLP with a narrow layer in the middle which is called a bottle-neck. Context-Dependent Deep Neural Network Hidden Markov Model (CD-DNN-HMM) [16, 17] is a hybrid system, which uses CD model for Large-Vocabulary Speech Recognition (LVSR) while (DNN-HMM) uses a layer-by-layer mechanism that provides well-extracted features that derive powerful noise-resistant features.

2.1.1.2 Feature Moment Normalization

Feature moment normalization methods is in normalizing the statistical moments of speech features. The Cepstral Mean Normalization (CMN) [18] is an example of Feature moment normalization methods, which is used in normalizing the first order statistical moments, Likewise, The Cepstral Mean and Variance Normalization (CMVN) [19] that is used in normalizing the second-order statistical moments. Lastly, higher order statistical moments normalizing using Histogram EQualization (HEQ) [20].

2.1.1.3 Feature Compensation

Feature compensation is used to remove the noise effect from the speech features. For example, Spectral Subtraction, which assumes that the noise and undistorted speech are uncorrelated and the noise characteristics change slowly compared to those of the speech waveform[21]. Thus, the noise can be suppressed by estimating the noise spectrum in the non-speech period. Wiener filtering is also used in removing noise by a real-valued gain function to remove the noise effects.

2.1.2 Model-space Approaches

In the Model-space approaches the acoustic model parameters are adjusted to decrease the noise effect. Although sometimes these approaches achieve a higher accuracy better than the Feature-space approaches but they require a higher computational time. This approach is divided into two sub-categories. The first category is the general adaptation, which adapts the mismatch between training and testing parameters by using general transformations techniques to adapt the acoustic

model factors. The second category is noise-specific methods, which compensates the model parameter as in general adaptation method by determining the nature of the distortions that is caused by the noise.

2.2 Automatic Speech Recognition System Architecture

As stated before, the main task of Automatic Speech Recognition (ASR) system is to convert an input of acoustic speech to a string of recognized words. The effect of the noisy channel on the acoustic waveform is reflected on the recognition accuracy. The Noise-Robust system architecture aims to find the best match between the noisy acoustic wave and the generated model from the previously trained speech data. There are two problems that face any ASR system. Firstly, the speech waveform is variable and varies from person to another. So, it is hard to exactly match any trained model with the uttered sentence. Secondly, in the case of huge language sentences, an efficient algorithm is needed to match the uttered sentence with a huge dataset.

In the system architecture, the Hidden Markov Model (HMM) is usually used. It is based on a probabilistic or Bayesian model, which is used in calculating the probability for a sequence of events. Due to the temporal characteristic of the speech waveform, The HMM is an effective tool to describe the changing of speech signals over the time. It models each word in the terms of consecutive of stochastic process. The ASR system architecture [22] is divided into two stage training stage and testing stage. The both stages are explicated as the follows:

2.2.1 Training Stage

Figure (2-1) demonstrate the block diagram of the training stage for ASR system. This stage targets to construct a stochastic models represents for each word. As shown in the block diagram, the speech training dataset that contains a set of speech waveform sentences is required. Moreover, the related lexicon dictionary that contains the phone structure for each word in the training set is also needed. All the vocabularies in the training dataset must recover all vocabularies that is needed to be recognized in the testing stage.

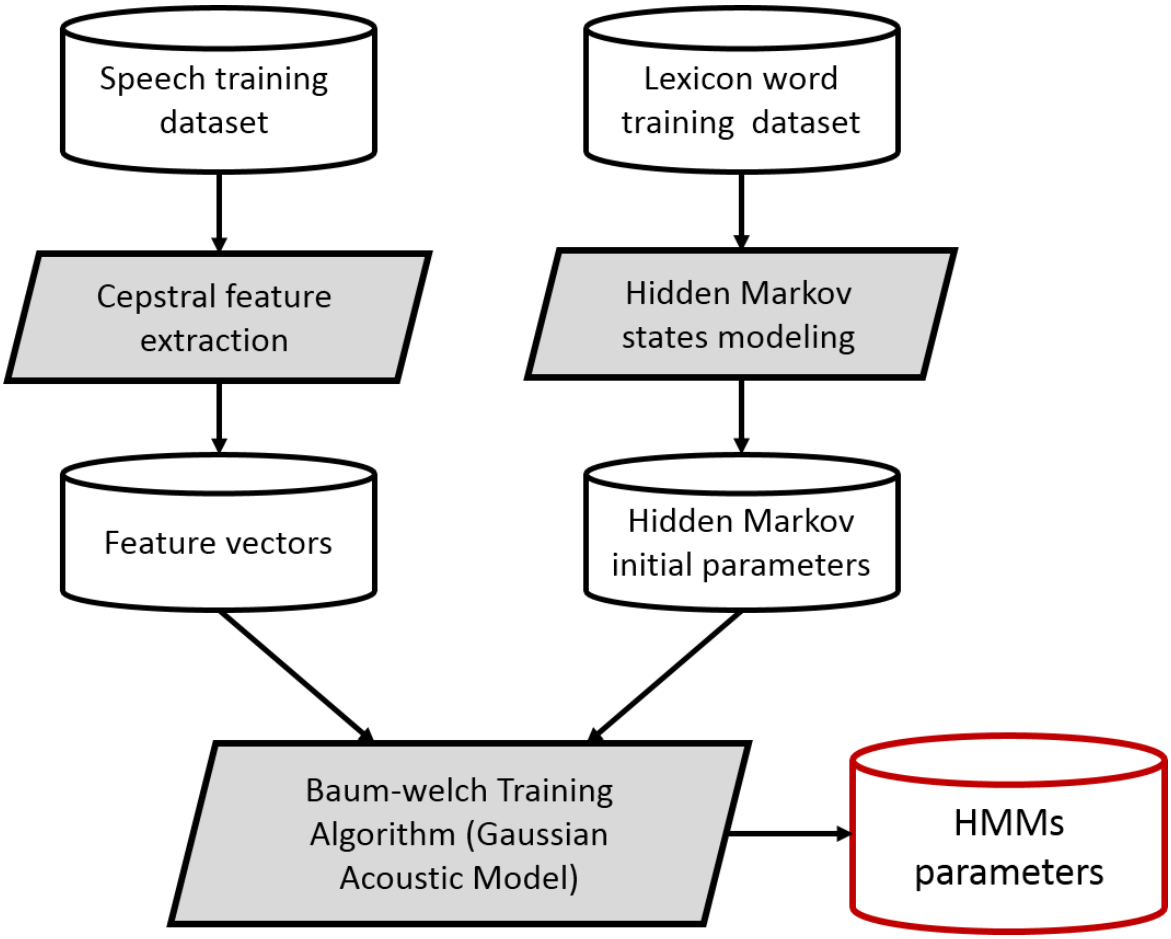


Figure 2-1: Block diagram of training stage for ASR system

2.2.1.1 Cepstral Feature Extraction

Cepstral feature extraction is the first process in the training stage [22]. In this section, each sentence in the training dataset is divided into short overlapped frames. Each frame is then converted to reduced feature vector o_i using one of the feature extraction techniques which will be explained in Section 2.3, the acoustic waveform feature input $O = [o_1, o_2, o_3, ..., o_T]$ can be treated as a sequence of individual symbols or observation where o_t represents temporally consecutive slices of input feature vector for a set of vocabulary words w_i.

2.2.1.2 Hidden Markov states modeling

To build the Markov states $Q = q_1, q_2, ..., q_t$ for each word, the one-to-one model which is also called left-to-right model or Bakis network is the most appropriate model which can describe the temporal consequence characteristic of the speech waveform. Each Markov state represents the phonemes of each word which is obtained from the related lexicon dictionary. For example for the word "Five" the number of phones are three (f – ay - v). But since each phone is not homogenous, it is modeled by three sub phones. Thus, the number of Markov states for the word "Five" will be nine and the sequence of phones. As in Figure (2-2).

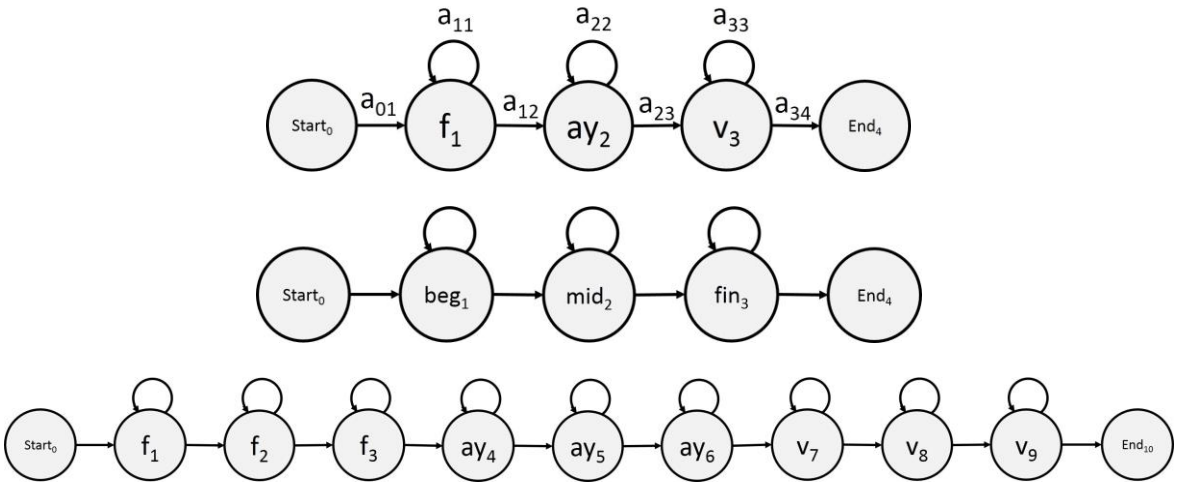


Figure 2-2: Markov states modeling for word "FIVE"

2.2.1.3 Baum-welch Training Algorithm

After generating the Markov states for each word, the extracted observations need to be classified to each sequence state. The observations vectors are modeled into a set of multivariate Gaussian models using a Gaussian Mixture Model (GMM) [23] technique. The GMM calculates the observation likelihoods $B_i = b_i(o_t)$ also called emission probabilities, each expressing the probability of an observation o_t.

$$b_j(o_t) = \sum_{\hat{m}=1}^{\hat{M}} \tilde{c}_{j\hat{m}} N(o_t; \mu_{j\hat{m}}, \Sigma_{j\hat{m}}) \qquad 2\text{-}1$$

where $\hat{M}$ is the number of the mixture parameters, $\tilde{c}_{j\hat{m}}$ is the weight of $\hat{m}^{th}$ parameter and $N(.; \mu, \Sigma)$ is multivariate Gaussian with mean vector μ and covariance matrix Σ as in.

$$N(o;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}e^{-\frac{1}{2}(o-\mu)^T\Sigma^{-1}(o-\mu)} \qquad 2\text{-}2$$

where n is the dimension of o. Since each mixture component can be considered as individual form of each sub-state, each observation likelihood can be considered as a single component Gaussian as the following:

$$b_j(o_t)=\frac{1}{\sqrt{(2\pi)^n\,|\Sigma_j|}}e^{-\frac{1}{2}(o_t-\mu_j)^T\Sigma^{-1}(o_t-\mu_j)} \qquad 2\text{-}3$$

Since the observation vectors underlay each state sequence is unknown, each observation vector is assigned to all states in proportion to the probability of the model being in that state. Therefore, $L_j(t)$ represents the probability of being in state j at time t. The maximum average likelihood μ_j and the covariance matrix Σ_j are calculated as the following:

$$\hat{\mu}_j=\frac{\sum_{t=1}^{T}L_i(t)o_t}{\sum_{t=1}^{T}L_i(t)} \qquad 2\text{-}4$$

$$\hat{\Sigma}_j=\frac{\sum_{t=1}^{T}L_i(t)(o_t-\mu_j)(o_t-\mu_j)^T}{\sum_{t=1}^{T}L_i(t)} \qquad 2\text{-}5$$

In the next step, the obtained likelihood value from the Gaussian Acoustic Modeling is formed by using Forward-Backward training algorithm as shown in Figure (2-3), which is called Baum-welch [24-26] training algorithm. The forward probability $\alpha_j(t)$ for the some model M with N states at $1<j<N$ and $1<t<T$ is calculated from the forward recursion process as:

$$\alpha_j(t)=P(o_1,o_2...o_t,q_t=j\,|\,M) \qquad 2\text{-}6$$

$$\alpha_j(t) = \left[\sum_{i=2}^{N=1} \alpha_i(t-1)a_{ij}\right] b_j(o_t) \qquad 2\text{-}7$$

The initial condition is assumed by setting $\alpha_1(1) = 1$

$$\alpha_j(1) = a_{1j} b_j(o_t) \qquad 2\text{-}8$$

For $1 < j < N$ the final condition is

$$P(O \mid M) = \alpha_N(T) = \sum_{i=2}^{N=1} \alpha_i(T) a_{iN} \qquad 2\text{-}9$$

Similarly, the backward probability $\beta_i(t)$ can also be calculated by recursion function for $1 < i < N$ and $T > t \geq 1$

$$\beta_j(t) = P(o_{t+1}, o_2 \ldots o_T, q_t = j \mid M) \qquad 2\text{-}10$$

$$\beta_j(t) = \sum_{j=2}^{N=1} a_{ij} b_j(o_{t+1}) \beta_j(t+1) \qquad 2\text{-}11$$

The initial conditions is assumed by setting $\beta_i(t) = a_{iN}$ and the final condition for $1 < i < N$ is:

$$P(O \mid M) = \beta_1(1) = \sum_{j=2}^{N=1} a_{1j} b_j(o_1) \beta_j(1) \qquad 2\text{-}12$$

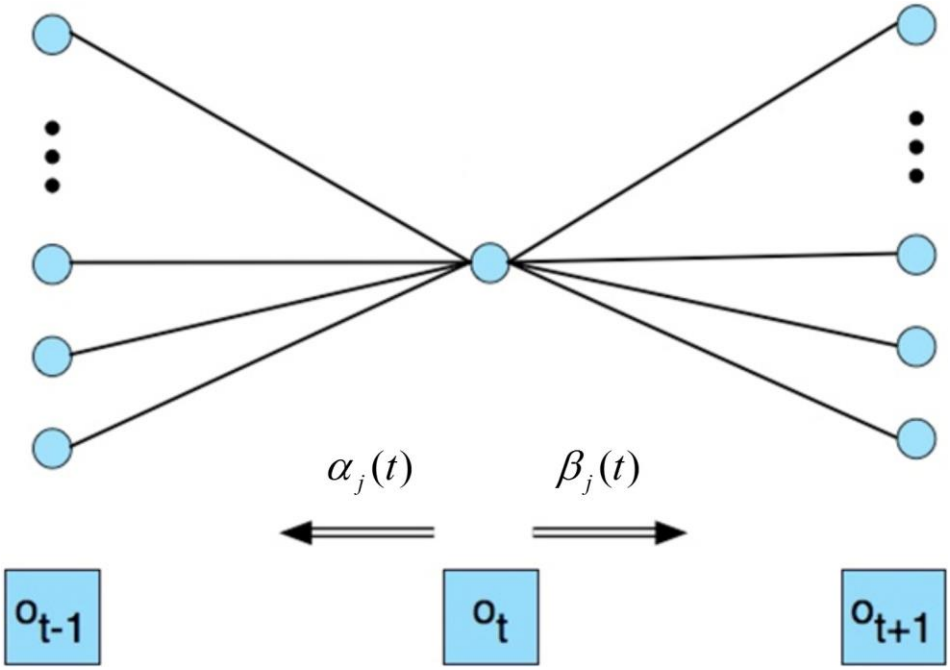


Figure 2-3: Baum-welch Forward-Backward algorithm

The total probability can be calculated from forward or backward probabilities as the following:

$$P = P(O \mid M) = \alpha_N(T) = \beta_1(1) \qquad 2\text{-}13$$

The forward probability represents the joint probability while the backward probability represents the conditional probability. Therefore, the probability of each state occupation is calculated by the product of two probabilities as in the following:

$$P(O, q_t = j \mid M) = \alpha_j(t)\beta_j(t) \qquad 2\text{-}14$$

$$L_j(t) = P(q_t = j, O \mid M) = \frac{P(O, q_t = j \mid M)}{P(O \mid M)} \qquad 2\text{-}15$$

$$L_j(t) = \frac{\alpha_j(t)\beta_j(t)}{P} \qquad 2\text{-}16$$

The transition probability can be re-estimated as the follows:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_i(t) a_{ij} b_j(o_{t+1}) \beta_j(t+1)}{\sum_{t=1}^{T} \alpha_i(t) \beta_i(t)} \qquad 2\text{-}17$$

where $1 < i < N$ and $1 < j < N$. The transition probability for entry state can be re-estimated by:

$$\hat{a}_{1j} = \frac{1}{p} \alpha_i(1) \beta_j(1) \qquad 2\text{-}18$$

The transition from the emitting state to final non-emitting state can be re-estimated by:

$$\hat{a}_{iN} = \frac{\alpha_i(T) \beta_j(T)}{\sum_{t=1}^{T} \alpha_i(t) \beta_i(t)} \qquad 2\text{-}19$$

The re-estimated transition probabilities, mean and covariance for each state are saved in database which will be used later in the testing stage.

2.2.2 Testing Stage

Figure (2-4) illustrates the Block diagram of testing stage for ASR system. In this stage, the uttered sentence is also divided into small overlapped frames and the cepstral features are extracted using the same feature extraction technique in the training stage. The constructed Markov models from training stage and N-gram language Model is used in this stage. The N-gram language Model is a dictionary contains a lattice structure that describes the grammar components which formulates the sentence. Using language Model can be a good benefit. If there are grammar relation between each word in the sentence. On the other hand, it can have a negative effect on speech recognition systems. If there are no grammar relation between each word in the sentence. For example, recognizing the telephone number digits.

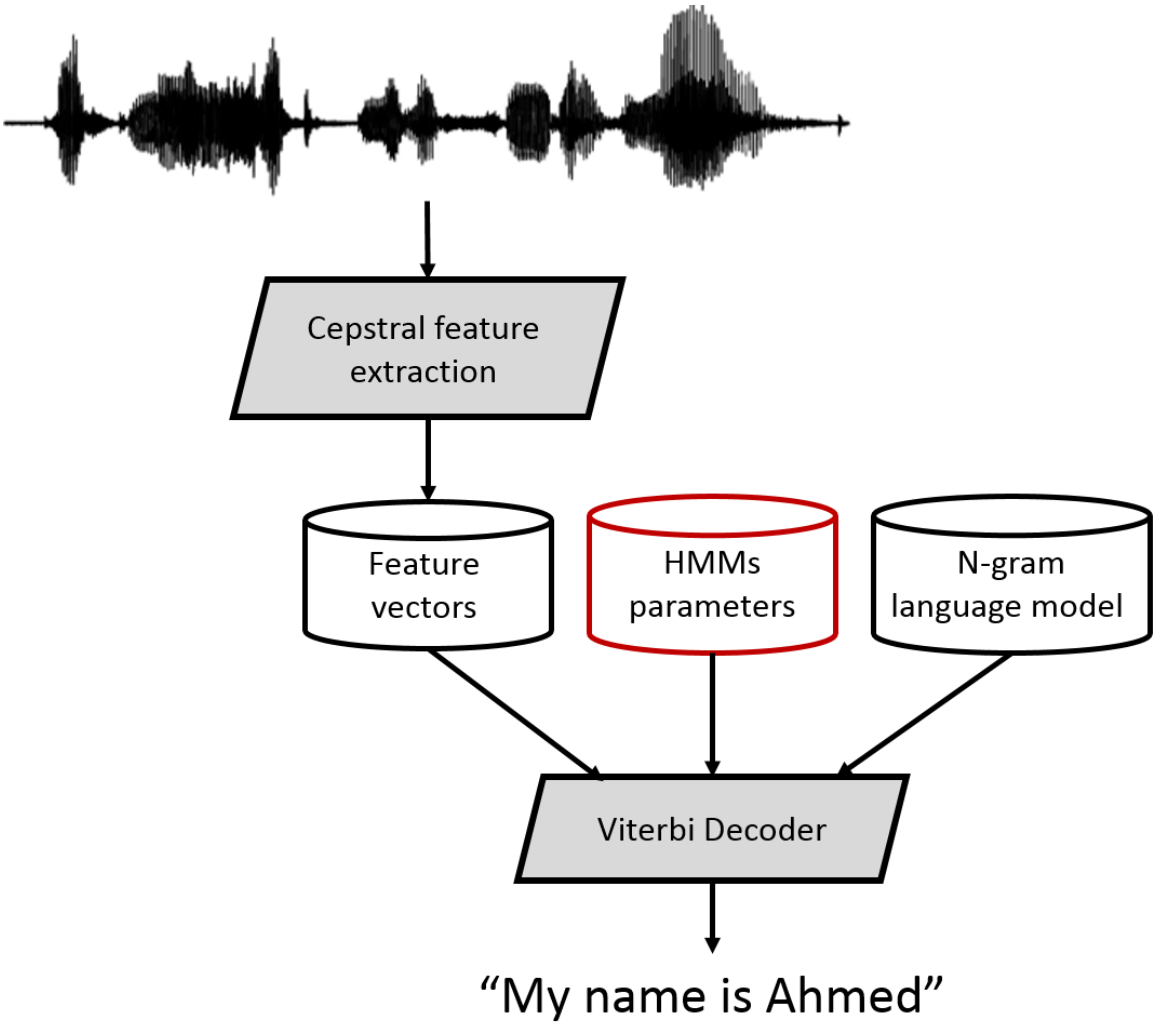


Figure 2-4: Block diagram of testing stage for ASR system

2.2.2.1 Decoding: The Viterbi Algorithm

In this stage, the Viterbi algorithm[27] is used to match speech features to one of the generated HMMs from the training stage in order to recognize each uttered word. The Viterbi algorithm uses the dynamic programming trellis algorithm. As shown in Figure (2-5), each Viterbi trellis element $v_t(j)$ represents the probability of the state j after t numbers of given observations.as shown in the following equation:

$$v_t(j) = \max_{q_0, q_1, \ldots, q_{t-1}} P(q_0, q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j \mid \lambda) \qquad 2\text{-}20$$

The maximum over all possible previous state sequences ($\max_{q_0,q_1,\cdots q_{t-1}}$) represents the most probable path. The Viterbi uses a recursive algorithm to calculate the probability of being in every state at time $t-1$.The Viterbi probability is computed by choosing the most probable path that leads to the current Viterbi element as shown in the following equation.

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \qquad 2\text{-}21$$

where $v_{t-1}(i)$ the previous Viterbi path probability, a_{ij} transition probability and $b_j(o_t)$ is the state observation likelihood.

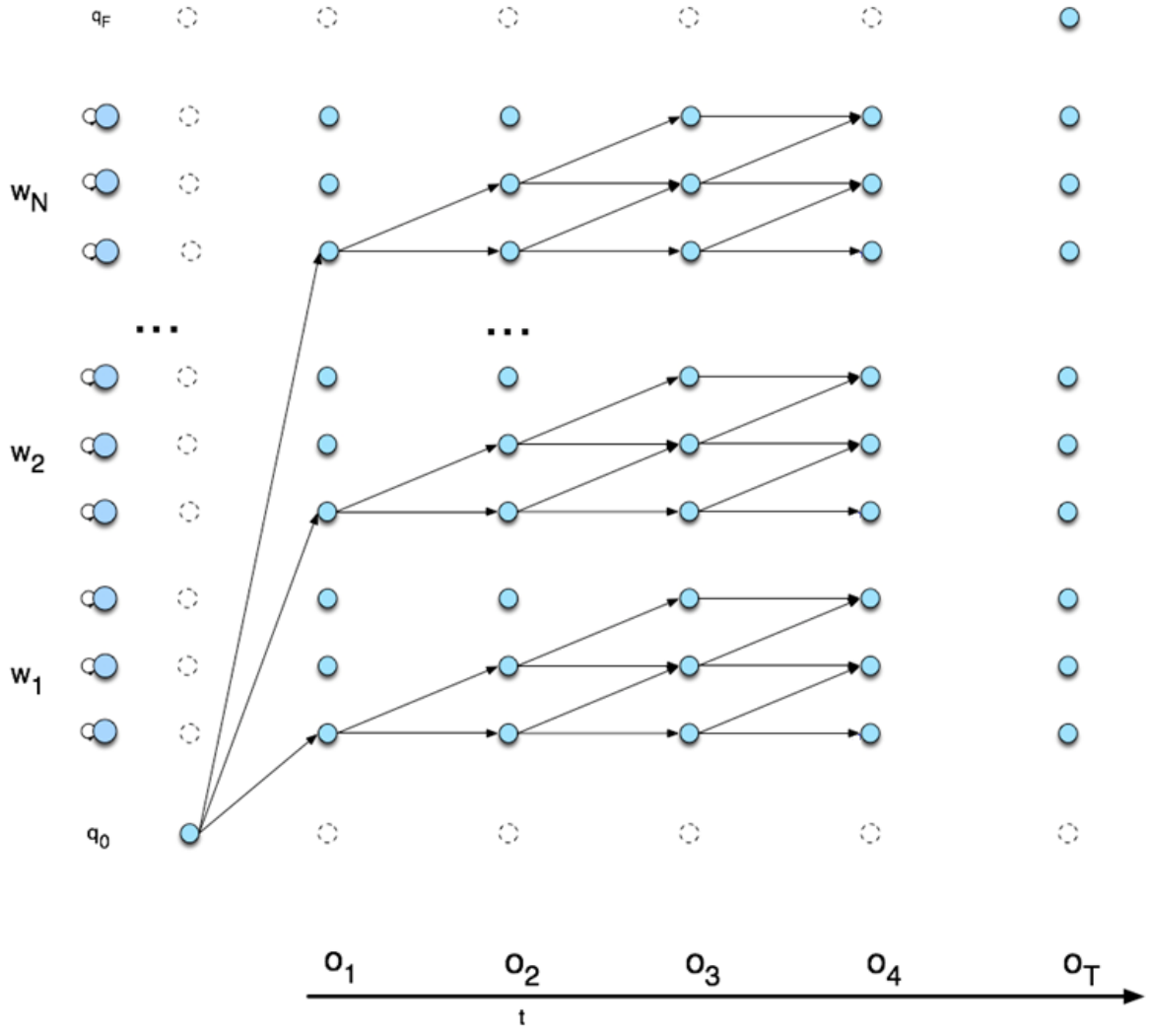


Figure 2-5: Viteribi decoder

2.3 Feature Extraction Techniques

Feature extraction is the first stage of speech recognition process. It converts the speech waveform data into reduced feature vectors by retaining the discriminative and non-redundant information in the speech data. These features represent the main characteristics of each word. There are a several feature extraction method are explicated in the following:

2.3.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a feature extraction model based on human auditory perception system [3]. It is one of the most popular speech feature extraction methods. The MFCC system diagram is explained as the Figure (2-6):

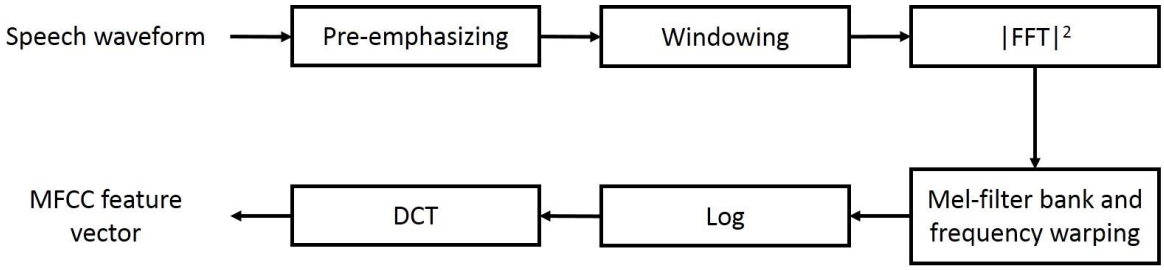


Figure 2-6: Block diagram of MFCC system

a) Pre-emphasizing

Most of the speech energy is concentrated in low frequencies more than middle and high frequency [28]. Pre-emphasizing is a first-order high-pass filter. It is applied to boost the spectrum values in high frequency.

$$y[n] = s[n] - \alpha s[n-1] \qquad 2\text{-}22$$

where $s[n]$ is the input signal, $s[n-1]$ previous sample and α is the filter coefficient in the range of $0.9 \le \alpha \le 1$ and the filter frequency response is shown in Figure (2-7).

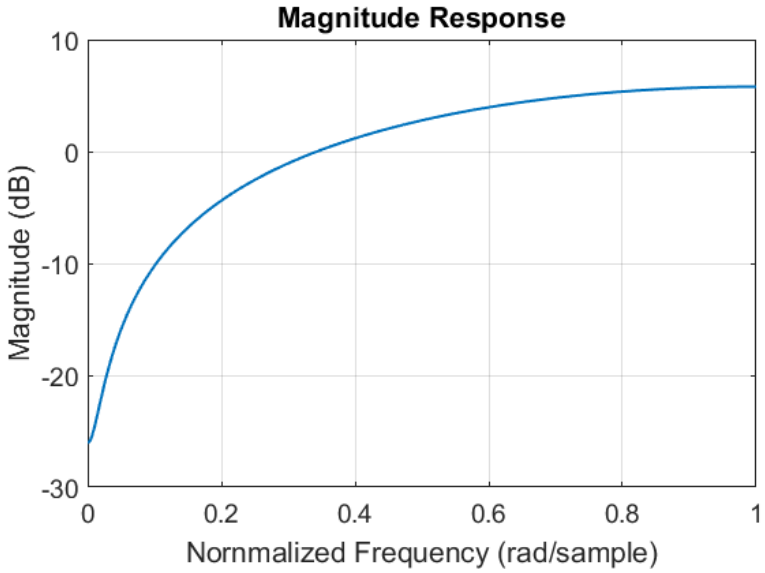


Figure 2-7: Magnitude response of pre-emphasize filter

b) Windowing

The speech waveform is quasi-stationary, therefore it is processed into short overlapped frames. Each frame width is between 20 to 30 ms and frame shift 10 ms as shown in Figure (2-8).

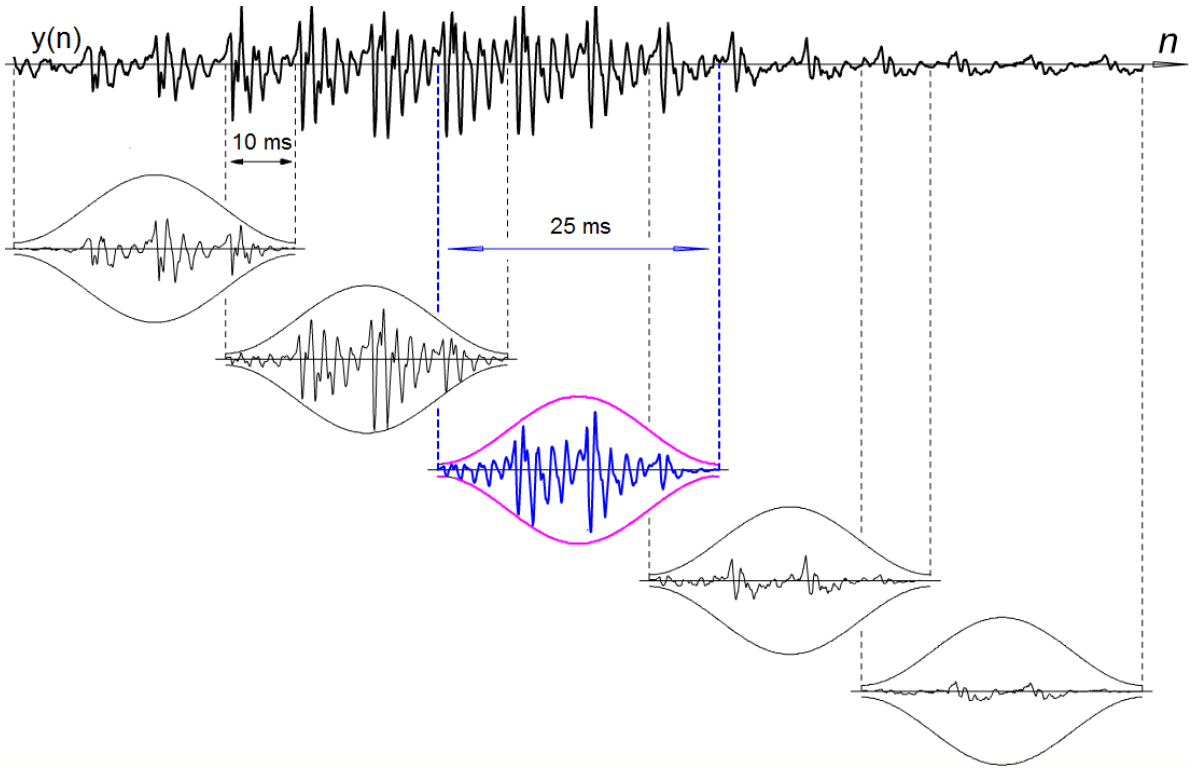


Figure 2-8: Windowing speech wave form

.according to the most literature. Then each frame is multiplied by Hamming window, which removes the sharp edge at the window boundaries.

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{J}\right) \qquad 2\text{-}23$$

where J is the total number of samples and n is integer number $0 \le n \le J-1$.

c) Power Spectral Density (PSD)

The Power Spectral Density (PSD) $|Y(k)|^2$ of each frame is then calculated, where it is referred to as the magnitude square of Fast Fourier Transform (FFT).

$$|Y(k)|^2 = \left|\sum_{n=1}^{N} y(n)\exp\left(-j\frac{2\pi nk}{N}\right)\right|^2 \qquad 2\text{-}24$$

where k is a frequency index, N is the total samples and n is a time index.

d) Mel-filter Banks and Frequency Warping

The human auditory system is not equally sensitive to all frequency values. It is sensitive to the low frequencies more than the high frequencies. The relation is called Mel-scale [29]. It is almost linear less than in the frequencies values that less than 1 kHz while it is wrapped to log scale in the frequencies values that more than 1 kHz as the following equation:

$$f_{mel} = 2595\log_{10}(1+\frac{f}{700}) \qquad 2\text{-}25$$

In MFCC system human auditory system is modeled as a triangular overlapped filters warped to Mel-scale as shown in the following Figures (2-9) and (2-10)

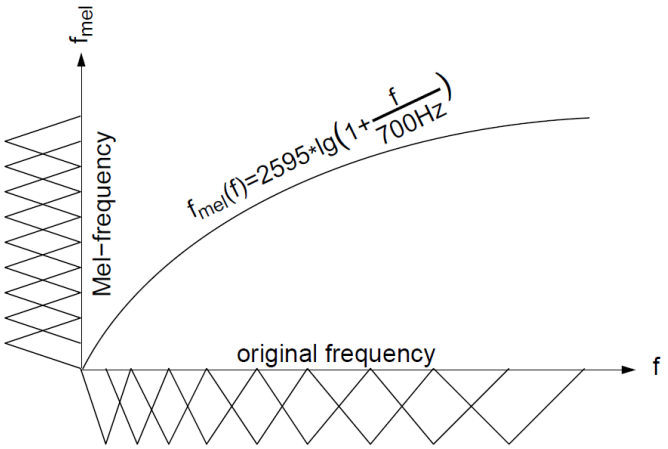


Figure 2-9: Frequency domain Mel-frequency scale relation

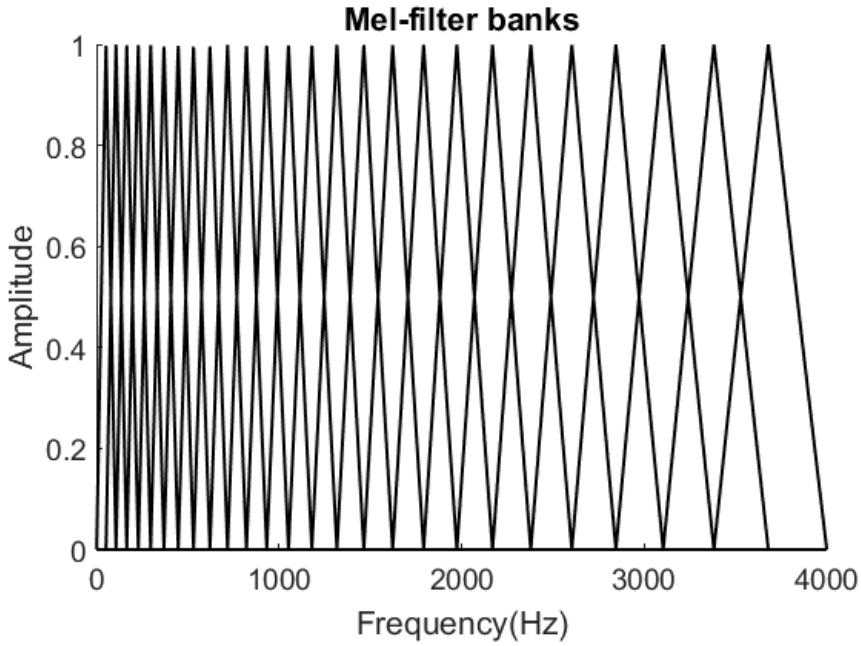


Figure 2-10: Triangular Mel-filter banks

Then, the power spectrum is convolved by a set of warped triangular filter banks and the resulting value of each filter bank $E(l)$ is calculated as shown in the following equation:

$$E(l) = \sum_{k=1}^{K/2} |Y(k)|^2 \Gamma_l(k) \qquad 2\text{-}26$$

where l is the filter index, $\Gamma_l(k)$ is the function of each filter in frequency domain and K is spectrum resolution.

e) Log Discrete Cosine Transform (DCT)

The cepstrum of the speech wave frame is calculated by applying the log operator on equation (2-27). Then the MFCC feature vector c_i is generated by the calculating the Discrete Cosine Transform (DCT).

$$c_i' = \sqrt{\frac{2}{L}} \sum_{m=1}^{L} \log\left(E(l)\right) \cos\left(\frac{\pi i}{L}(m-0.5)\right) \qquad 2\text{-}27$$

where L is total number MFCC features and i feature vector index.

2.3.2 RelAtive SpecTrAl Perceptual Linear Predictive (RASTA-PLP)

PLP is another approach or method that models the human auditory system. It is based on three techniques which are derived from the physiology of hearing. The first technique is the critical-band spectral resolution, the second technique is the equal-loudness curve, and the last technique is the intensity-loudness power law. RASTA process is a noise filtering technique that is combined with the PLP system, which is illustrated in the block diagram in Figure (2-11). The preprocessing of RASTA-PLP system is similar to MFCC system except using the pre-emphasize process. The speech waveform is divided into overlapped frames and multiplied by hamming window, The PSD of each frame is calculated. The rest processes are different and explicated as the follows:

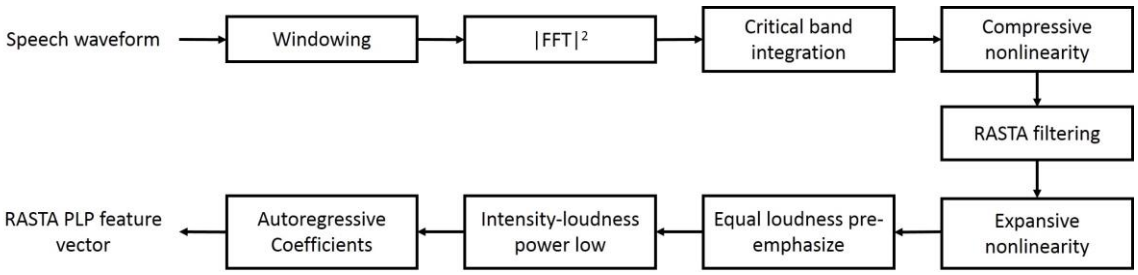


Figure 2-11: Block digram of RASTA-PLP system

a) Critical band Integration

The spectrum is warped to Bark frequency scale Ω by the following equation:

$$\Omega(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right] \qquad 2\text{-}28$$

where ω is an angular frequency in rad/s, Then warped power spectrum is convolved with a critical band masking curve $\Psi(\Omega)$ each making curve represents the Bark filter bank which is given by

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega-0.5)} & -1.3 \le \Omega \le -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-0.1(\Omega-0.5)} & 0.5 \le \Omega \le 2.5 \\ 0 & \Omega > 2.5 \end{cases} \qquad 2\text{-}29$$

The discrete convolution of filter function $\Psi(\Omega)$ with power spectrum function $|Y(\omega)|^2$ is shown in Equation (3-30) with 0.994-Bark steps. The generated Bark filter banks for spectrum resolution 256 is shown in Figure (2-12).

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} |Y(\Omega - \Omega_i)|^2 \Psi(\Omega) \qquad 2\text{-}30$$

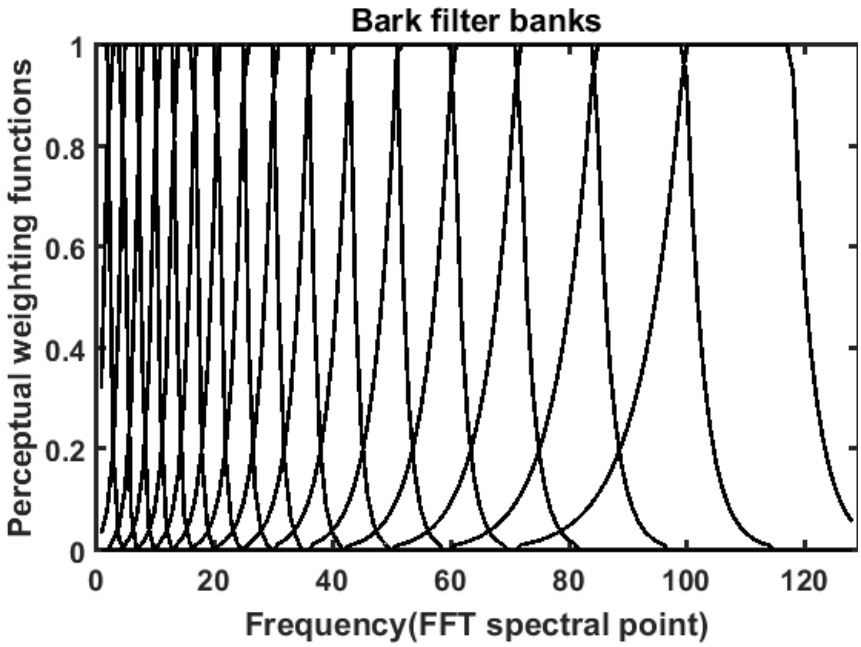


Figure 2-12: Bark filter banks

b) RASTA filtering

The human hearing is insensitive to slow change in frequency characteristic of the communication characteristic. Each frequency channel that is produced from the critical band integration stage is filtered using a band pass filter with a sharp spectral zero at zero frequency in order to suppress the steady background noise. This filtering technique is applied in three steps the first step is compressive nonlinearity by calculating the log scale of the spectrum band. The second step is applying a RASTA band pass filter function as in Equation (2-31). Last but not least, expansive nonlinearity is applying by calculating the inverse log or the exponential on the filterd spectrum.

$$H(z) = 0.1z^{4}\,\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \qquad 2\text{-}31$$

c) Equal loudness pre-emphasize

The human hearing is not equally sensitive at different frequencies. Therefore the pre-emphasize equal-loudness relation $E(\omega)$ is applied.

$$\mathrm{E}[\Omega(\omega)] = E_{pre}(\omega)\Theta[\Omega(\omega)] \qquad 2\text{-}32$$

$$E_{pre}(\omega) = \frac{(\omega^2 + 56.8\times10^6)\omega^4}{(\omega^2 + 6.3\times10^6)^2 \times (\omega^2 + 0.38\times10^9)} \qquad 2\text{-}33$$

d) Intensity-loudness power low

This operation simulates the nonlinear relation between the intensity of sound and its perceived loudness by applying cubic-root amplitude compression as the following:

$$\Phi(\Omega) = \left(\mathrm{E}(\Omega)\right)^{0.33} \qquad 2\text{-}34$$

e) Autoregressive Coefficients

The last stage is approximating the spectrum using all-pole spectral modeling technique. In this technique, the Inverse Fast Fourier Transform (IFFT) is applied to $\Phi(\Omega)$. Then, cepstral coefficients are calculated from autoregressive coefficients, which are generated using an all-pole modeling technique.

2.3.3 Gammatone Frequency Cepstral Coefficients (GFCC)

Figure (2-13) illustrates the block diagram of PNCC system. Likewise MFCC system, the pre-emphasizing is calculated for speech waveform. Next, the pre-emphasized speech is framed and each frame is multiplied by a Hamming window. Then, the PSD is calculated. In GFCC system, Gammatone filter banks were used to model the human auditory perception system instead of triangular Mel-filter banks that is used in MFCC system. Finally, the log and DCT processes are applied to obtain the GFCC fractures.

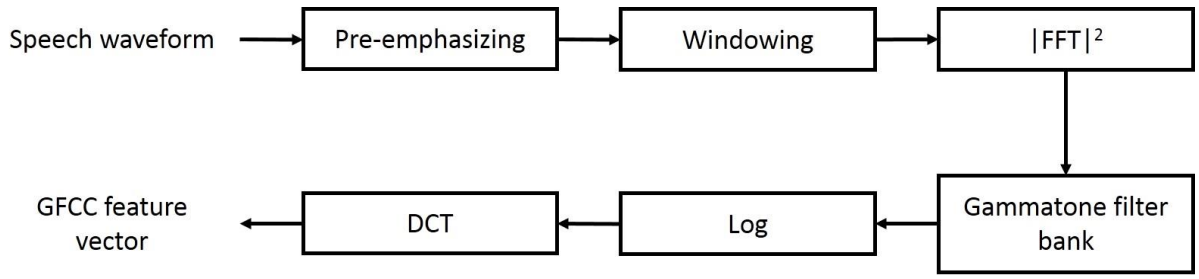


Figure 2-13: Block diagram of GFCC system

a) Gammatone filter banks

Figure (2-14) demonstrates the cochlea physiology in the human auditory system. At first, the mechanical force is produced from the perceived sound vibration. Since the structures within the cochlea are not rigid, the basilar membrane is flexible and bends in response to sound. The motion shape of the basilar membrane within the cochlea resembles a tone modulated with a gamma function in the time domain [30]. The stiffness of the membrane is reduced from base to apex [31]. The high-frequency sounds have higher energy. So, it vibrates stiffer part of the basilar membrane. On the other hand, The Lower-frequency sounds have lower energy and it vibrates the basilar membrane at the apex. Hence, the basilar membrane vibration is more sensitive to low frequencies more than the high frequencies. Then the produced vibration is analyzed by the by the auditory nerve fiber, which sends signals to the brain to recognize the perceived sound.

.

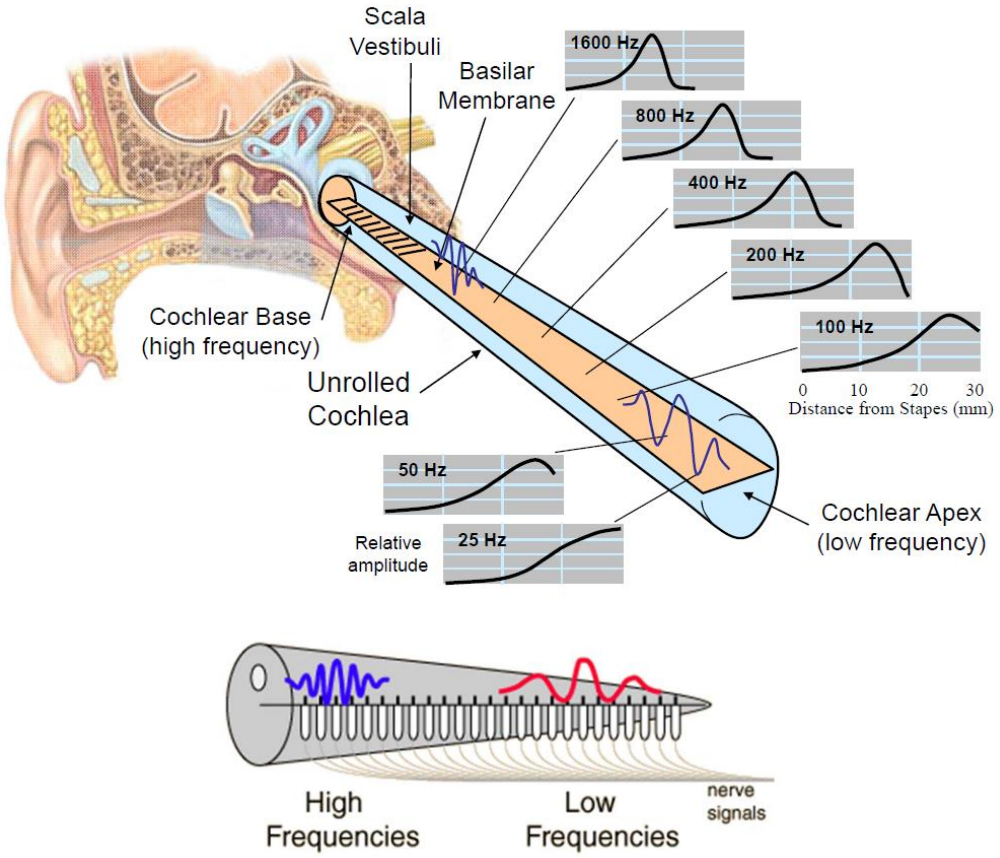


Figure 2-14: Physiology of the cochlea in the human auditory system

The Gammatone filter simulates the frequency domain of the impulse response of the basilar membrane output to the auditory nerve fiber [30, 32], which product of a gamma distribution and sinusoidal tone as in Equation (2-35).

$$g(t) = \hat{a}t^{g-1}e^{-2\pi\hat{b}(f_c)t}\cos(2\pi f_c t + \phi) \qquad 2\text{-}35$$

Where f_c is the center frequency, ϕ is the phase of the carrier, $\hat{a}$ is the amplitude, g is the filter's order, $\hat{b}$ is the filter's bandwidth, t and is time. The set of 25 Gammatone impulse responses from center frequencies 100 Hz to 4 KHz are generated as shown in Figure (2-16) and the frequency domain of each impulse responses is shown in Figure (2-16).

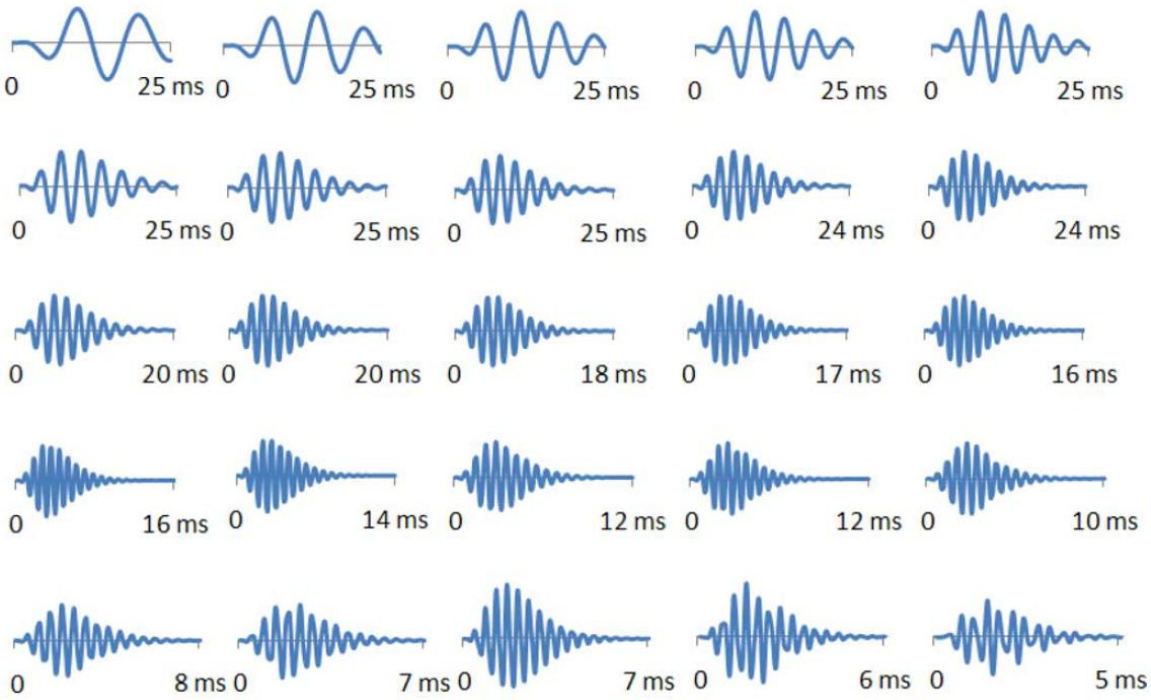


Figure 2-15: Set of 25 gammatone impulse response with center frequencies from 100 Hz to 4 KHz

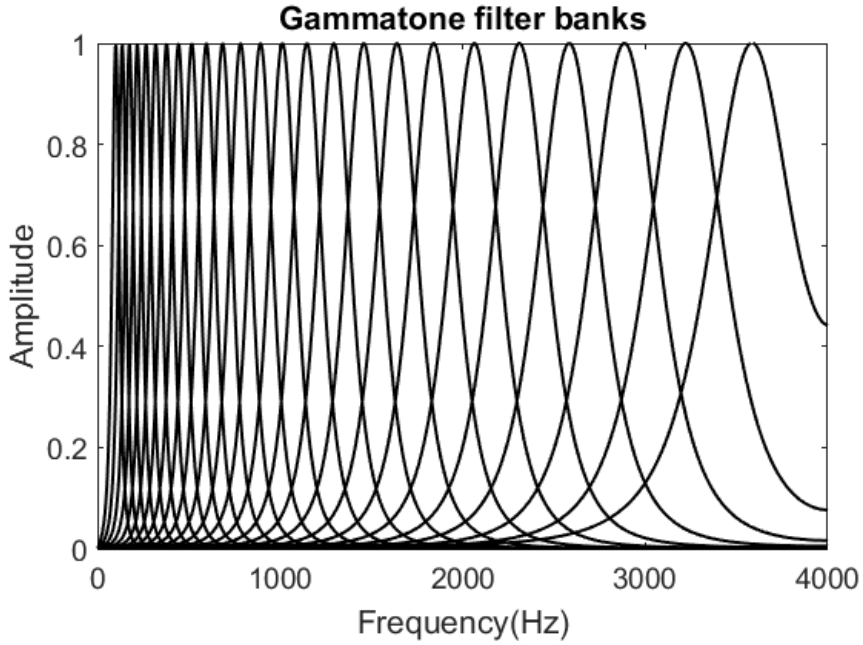


Figure 2-16: Gammatone filter banks

2.3.4 Power-Normalized Cepstral Coefficients (PNCC)

Power-Normalized Cepstral Coefficients is one of state-of-the-art robust feature extraction system. There are many versions for PNCC system [33, 34]. The latest version [12] is implemented in this research where some modifications are considered, for example, using "medium-time" processing within a duration of 50–120 ms to analyze the effect of the noise, an "asymmetric nonlinear filtering" to determine the background noise level and using "power-law nonlinearity" instead of "log nonlinearity".

The system is implemented on a 16 kHz speech dataset. The block diagram of the system is shown in Figure (2-17). The initial processing is as same as the previously mentioned systems. The pre-emphasis is applied to the speech waveform and then it is framed into overlapped frames. Each frame duration is 25.6 ms and the frame shift is 10 m. After that, each frame is multiplied by a Hamming window and the PSD is calculated with 1024 spectral resolution for each frame.

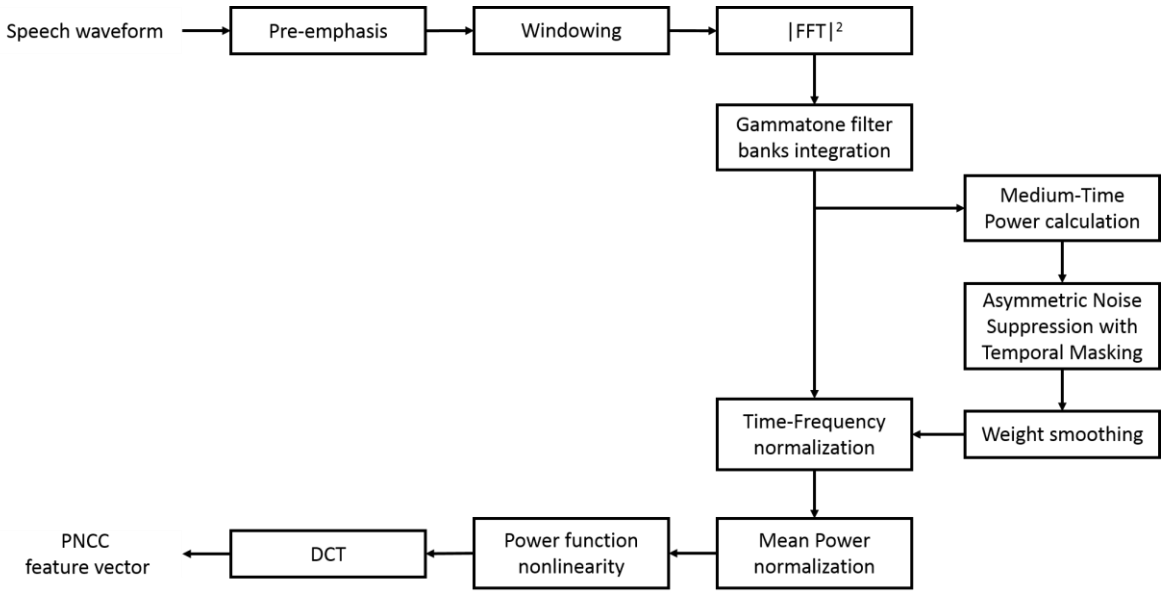


Figure 2-17: Block diagram of PNCC system

A set of 40 Gammatone filters are produced by calculating the frequency response of generated Gammatone frequencies from the range of 200 Hz to 8 kHz, which is shown in Figure (2-18). Each filter bank is normalized and the value is equal to zero if the filter bank is less than of 0.5 percent of the maximum value. The short-time spectral power is calculated using $P[m,l]$ as the following:

$$P[m,l] = \sum_{k=1}^{(K-1)/2} \left| Y[m,k] G_l(k) \right|^2 \qquad 2\text{-}36$$

where m represent the frame number, l is Gammatone channel index and K is the spectrum resolution.

a) Medium-Time Power Calculation

While the power that related with most background noise changes slower than the power that

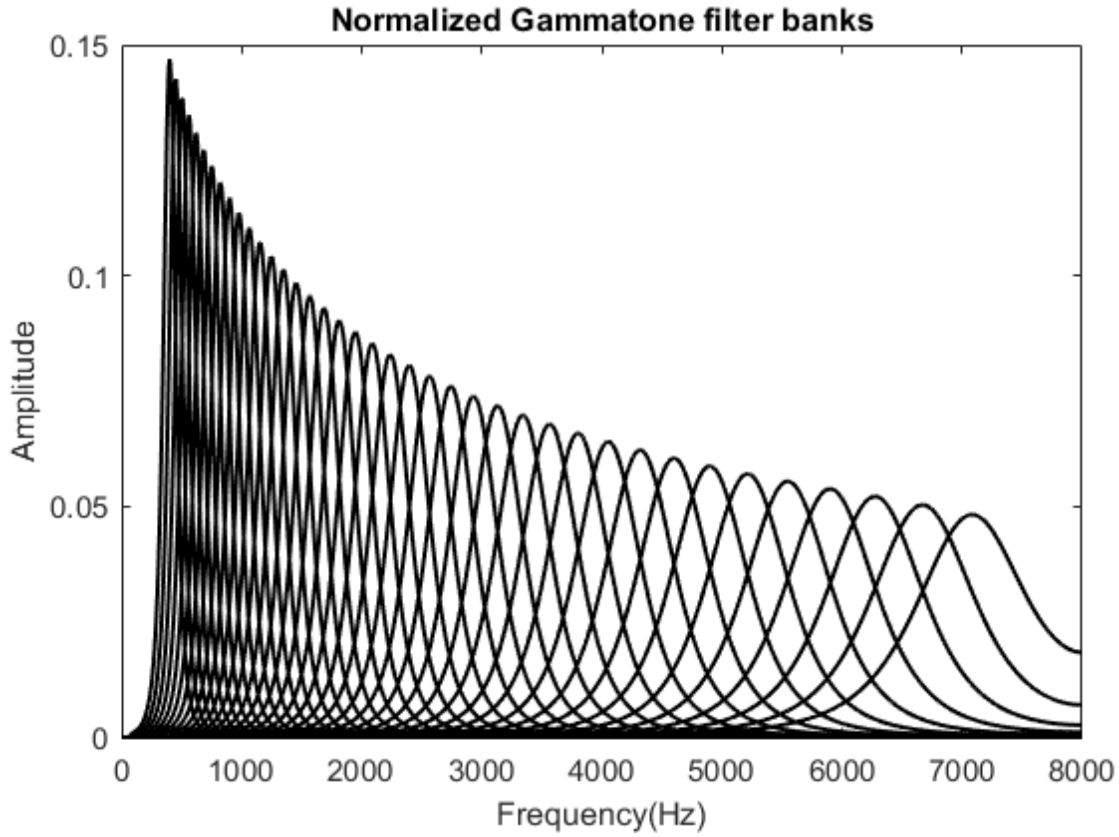


Figure 2-18: Normalized 40 Gammatone filter banks

associated with the speech. The medium-time power $\tilde{Q}[m,l]$ filtering technique is applied by computing the running average power $P[m,l]$ as in the equation:

$$\tilde{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m+M}^{m+M} P[m',l] \qquad 2\text{-}37$$

In PNCC system, the value of M is chosen to equal 2 because the higher value will blur the onsets and offsets of the frequency and that will degraded the system performance.

b) Asymmetric Noise Suppression with Temporal Masking

In this section, a nonlinear filter suppression technique is applied. This approach used in removing slowly varying back ground noise. The block diagram for Asymmetric Noise Suppression process with Temporal Masking is shown in Figure (2-19).

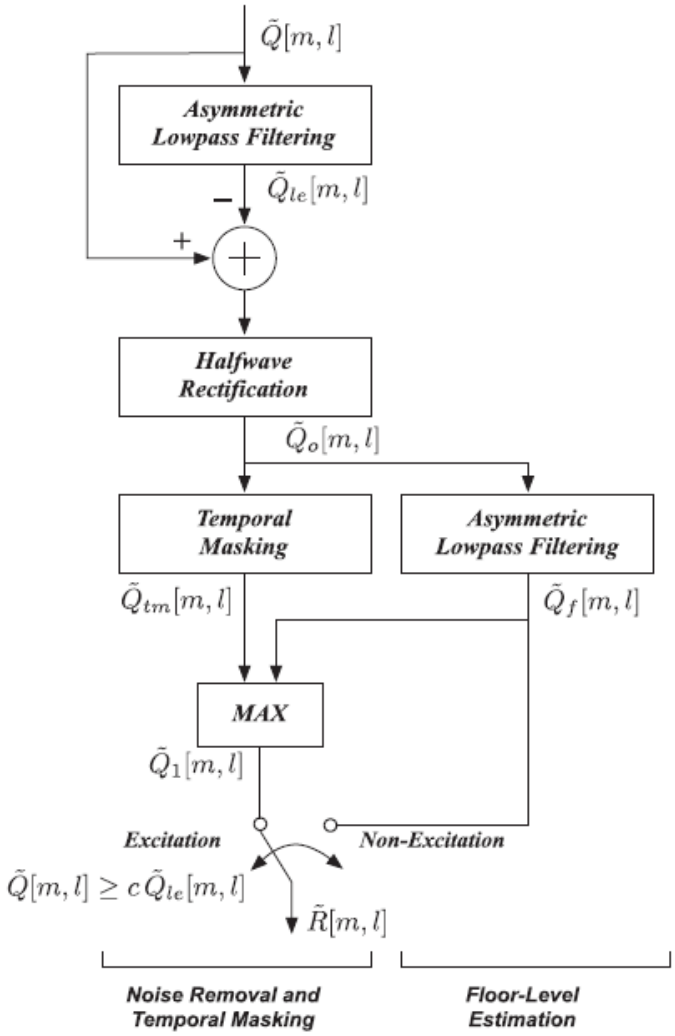


Figure 2-19: Block diagram of Asymmetric Noise Suppression with Temporal Masking

In the first stage, an asymmetric nonlinear filtering which is computed as shown in the following equation where $\tilde{Q}_{in}[m,l]$ and $\tilde{Q}_{out}[m,l]$ are an arbitrary input and output, respectively.

$$\tilde{Q}_{out}[m,l] = AF_{\lambda_a,\lambda_b}[\tilde{Q}_{in}[m,l]] \qquad 2\text{-}38$$

where λ_a and λ_b are constants between zero and one

$$\tilde{Q}_{out}[m,l] = \begin{cases} \lambda_a \tilde{Q}_{out}[m-1,l] + (1-\lambda_a)\tilde{Q}_{in}[m,l], \\ \quad if \tilde{Q}_{in}[m,l] \geq \tilde{Q}_{out}[m-1,l] \\ \lambda_b \tilde{Q}_{out}[m-1,l] + (1-\lambda_b)\tilde{Q}_{in}[m,l], \\ \quad if \tilde{Q}_{in}[m,l] < \tilde{Q}_{out}[m-1,l] \end{cases} \qquad 2\text{-}39$$

In the case of $1 > \lambda_a > \lambda_b > 0$, the filter output $\tilde{Q}_{out}[m,l]$ will lead to lower envelope of $\tilde{Q}_{in}[m,l]$. The slowly-varying lower envelope is assumed to equal medium-time noise level while the activity above this envelope is estimated to represent speech activity. Thus, subtracting this low-level envelope from the input $\tilde{Q}_{in}[m,l]$ will remove a slowly varying non-speech components. The values of constants λ_a and λ_b are chosen to equal to 0.999 and 0.5, respectively.

$$\tilde{Q}_{le}[m,l] = AF_{0.999,0.5}[\tilde{Q}[m,l]] \qquad 2\text{-}40$$

The input is high-pass filtered by subtracting $\tilde{Q}_{le}[m,l]$ from $\tilde{Q}[m,l]$, then using half-wave rectifier, the negative values of the produced signal is set to zero and the produced output value is $\tilde{Q}_0[m,l]$. After that, two process were applied to the output value. The first process, is an asymmetric non-linear low-pass filter that is applied again to the rectifier output to get the lower envelope $\tilde{Q}_f[m,l]$ that is obtained as the following:

$$\tilde{Q}_f[m,l] = AF_{0.999,0.5}[\tilde{Q}_0[m,l]] \qquad 2\text{-}41$$

The second process, is temporal masking which is demonstrated in Figure (2-20). It is based on the concept of the human auditory system appears to concentrate on the incoming power envelope more than the falling edge of that same power envelope [35, 36]. This process is shown in the block diagram in Figure (2-20).

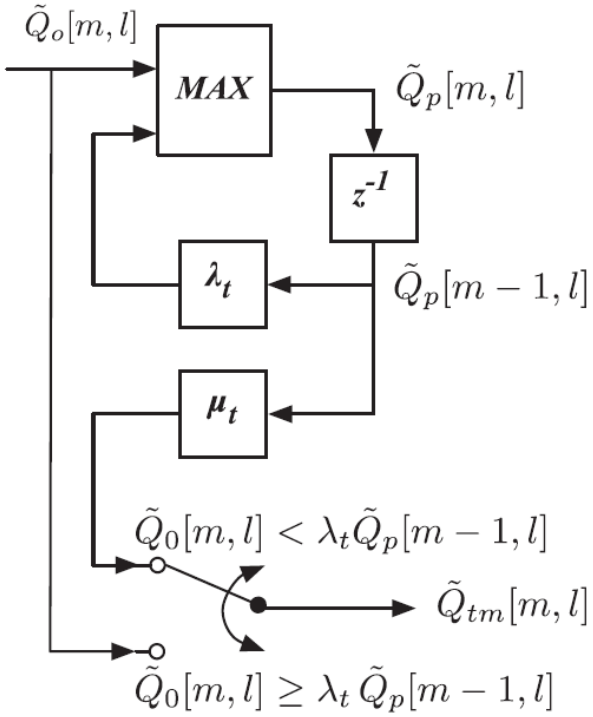


Figure 2-20: Block diagram of Temporal Masking

At first, the on-line peak power $\tilde{Q}_p[m,l]$ is obtained for each channel using the following equation:

$$\tilde{Q}_p[m,l] = \max(\lambda_t \tilde{Q}_p[m-1,l], \tilde{Q}_0[m,l]) \qquad 2\text{-}42$$

where λ_t is a forgetting factor which is equal to 0.85 and the Temporal masking for speech segments is determined using the following equation:

$$\tilde{Q}_{tm}[m,l] = \begin{cases} \tilde{Q}_0[m,l], & \tilde{Q}_0[m,l] \geq \lambda_t \tilde{Q}_p[m-1,l] \\ \mu_t \tilde{Q}_p[m-1,l], & \tilde{Q}_0[m,l] < \lambda_t \tilde{Q}_p[m-1,l] \end{cases} \qquad 2\text{-}43$$

where μ_t is a constant equal to 0.2. The output processing is then obtained from the lower envelope of the rectified signal $\tilde{Q}_f[m,l]$ and the temporal masking output $\tilde{Q}_{tm}[m,l]$ as shown in the following equation:

$$\tilde{Q}_1[m,l] = \max(\tilde{Q}_{tm}[m,l], \tilde{Q}_f[m,l]) \qquad 2\text{-}44$$

The last process, the low envelope of the rectified signal represents the low-pass filtered noise segment while the speech segment is not low pass filtered to avoid blurring the power coefficients of the speech. Thus, If $\tilde{Q}[m,l] \geq c\tilde{Q}_{le}[m,l]$ the value of $\tilde{R}[m,l]$ is equal to $\tilde{Q}_1[m,l]$ which determines the Excitation segment. Meanwhile, in the case of $\tilde{Q}[m,l] < c\tilde{Q}_{le}[m,l]$ the value of $\tilde{R}[m,l]$ is equal to $\tilde{Q}_f[m,l]$ which determines the Non-excitation segment. The constant c is fixed threshold equal to 2 which provides the best performance for the white noise.

c) Spectral Weight Smoothing

In this section, the asymmetric noise suppression is combined to temporal masking for each time frame and frequency value according to the transfer function $\tilde{R}[m,l]/\tilde{Q}[m,l]$. The transfer function is smoothed across frequency by calculating the running average over the channel index l of the ratio $\tilde{R}[m,l]/\tilde{Q}[m,l]$.

$$\tilde{S}[m,l] = \left(\frac{1}{l_2 - l_1 + 1} \sum_{l'=l_1}^{l_2} \frac{\tilde{R}[m,l']}{\tilde{Q}[m,l']} \right) \qquad 2\text{-}45$$

where $l_2 = \min(l+N, L)$ and $l_1 = \min(l-N, 1)$, the value of N is equal 4 for 40 numbers of L channels.

d) Time-Frequency Normalization

The Time-Frequency Normalization is obtained by modulated the time-averaged, frequency-averaged transfer function $\tilde{S}[m,l]$ to the short-time power $P[m,l]$ as in the following equation:

$$\tilde{T}[m,l] = P[m,l]\tilde{S}[m,l] \qquad 2\text{-}46$$

e) Mean Power Normalization

The human auditory processing contains an automatic gain control that decreases the impact of amplitude variation of the incoming acoustic wave. In the PNCC system, the power-law nonlinearity which is explicated below, causes the response of the processing to be affected by the variation in absolute power, although this effect is usually small. The Mean Power Normalization stage is applied to minimize the potential effect of amplitude scaling. The input power is normalized by dividing the received power by a running average of the total power. Firstly, the mean power estimate $\mu[m]$ is calculated from the following equation:

$$\mu[m] = \lambda_\mu \mu[m-1] + \frac{(1-\lambda_\mu)}{L} \sum_{l=0}^{L-1} \tilde{Q}[m,l] \qquad 2\text{-}47$$

where λ_m is a forgetting factor whose value is 0.999. Then the normalized power is computed directly from the running power estimate $\mu[m]$ as the following equation:

$$U[m,l] = k \frac{\tilde{T}[m,l]}{\mu[m]} \qquad 2\text{-}48$$

where k an arbitrary constant.

f) Power Function Nonlinearity

The relation between the nonlinear sound pressure level in decibels to the auditory-nerve firing rate is compressive [37, 38]. In addition to, the average auditory-nerve firing rate shows an overshoot at the beginning of an input signal. It has found experimentally in [39] that a power-law curve with the power of 1/15 for sound pressure presents a proper fit to the physiological data while optimizing recognition accuracy for the noisy speech.

$$V[m,l] = U[m,l]^{1/15} \qquad 2\text{-}49$$

2.3.5 Cepstral Mean Normalization (CMN)

CMN is applied to move all of the cepstral features to have a zero mean as shown in Figure (2-21). It is calculated in two steps. The first step is computing the mean of all cepstral feature

vectors along the frames as in Equation (2-50). Afterward, the computed mean is subtracted from all the cepstral vectors as in Equation (2-51).

$$\tilde{m}_i = \frac{1}{T}\sum_{m=1}^{T} c_i'(m) \qquad 2\text{-}50$$

$$c_n'(m) = c_i'(m) - \tilde{m}_i \qquad 2\text{-}51$$

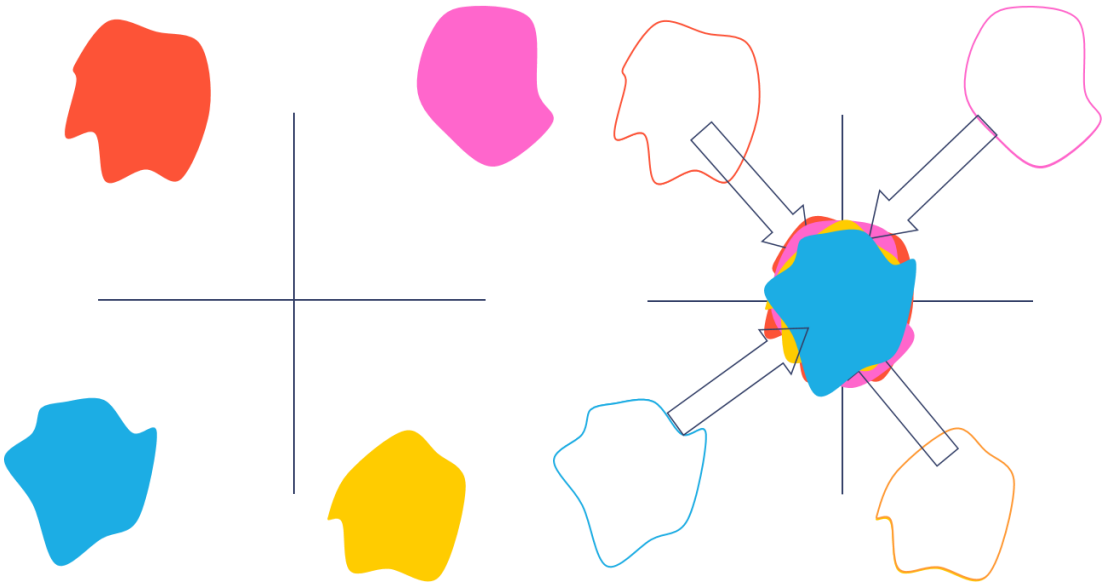


Figure 2-21: Moving cepstral features to have a zero mean

2.3.6 Energy Features

The energy feature is calculated by computing the summation of the power of the samples over the frame from the sample at time t_1 to the sample at time t_2 as in the following formula:

$$Energy = \sum_{n=t1}^{t2} s^2(n) \qquad 2\text{-}52$$

2.3.7 Dynamic Features Delta (Δ) and Delta-delta (ΔΔ)

The delta features (Δ) represents the velocity variation of the features over the frames as in Equation (2-53). The delta-delta features (ΔΔ) are the acceleration variation of the features over the frames and it is obtained from the velocity features as in Equation (2-54).

$$\Delta d(t) = \frac{c'(t+1) - c'(t-1)}{2} \qquad 2\text{-}53$$

$$\Delta\Delta d(t) = \frac{\Delta d(t+1) - \Delta d(t-1)}{2} \qquad 2\text{-}54$$

2.4 Performance Evaluation

The difficulty of evaluating the performance of the speech recognition system lies in the fact that the recognized words within the sentence can have the different length of the sequence of words within the reference sentence. The total number of words in the reference sentence is defined by symbol N while there are three numbers that represent the source of errors. The first number is substitution error S, which represents the number of correct words in the reference sentence that is substituted by the incorrect words. The second number is deletion error D, which defines the number of words which were existed in the reference sentence and removed in the recognized sentence. The third number is insertion error I, which is the number of words that were inserted to the recognized sentence and they were not existed in the reference sentence. The performance of the ASR system can be evaluated in two different terms. The first term is percentage Word Accuracy as shown in Equation (2-55).The second term is percentage Word Recognition Rate (WRR) or percentage Correct word which is similar to percentage Word Accuracy but the insertion error is ignored as in Equation (2-56).

$$Acc = \frac{N - S - D - I}{N} \qquad 2\text{-}55$$

$$WRR = \frac{N - S - D}{N} \qquad 2\text{-}56$$

CHAPTER 3

PROPOSED METHOD

3 Chapter THREE: Proposed Method

3.1 Proposed Method 1: Modified MFCC Technique

The first proposed method was developed by a desire to improve the extracted speech features in the presence of Additive White Gaussian Noise (AWGN). The system was designed to robust against the noise without loss of the performance in case of the undistorted speech waveform. The proposed system is based on MFCC feature extraction system as shown in block diagram in Figure (3-1).

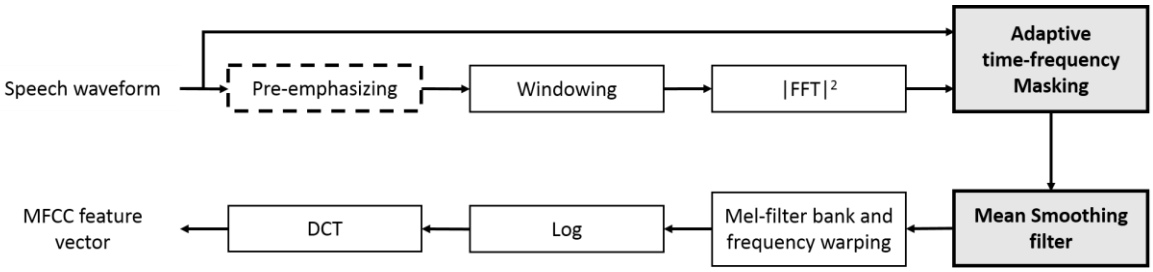


Figure 3-1: Block diagram of the proposed MFCC method

In the proposed method, the pre-emphasizing was removed to prevent boosting the noise amplitude at high frequency value. The uttered word is framed and multiplied by a hamming window. Then, the PSD of each frame is calculated and a time-frequency map for each word is constructed.

The adaptive mask is estimated for each word from the constructed time-frequency map. The block diagram of the proposed adaptive time-frequency masking is shown in Figure (3-2) and it is explicated as the following:

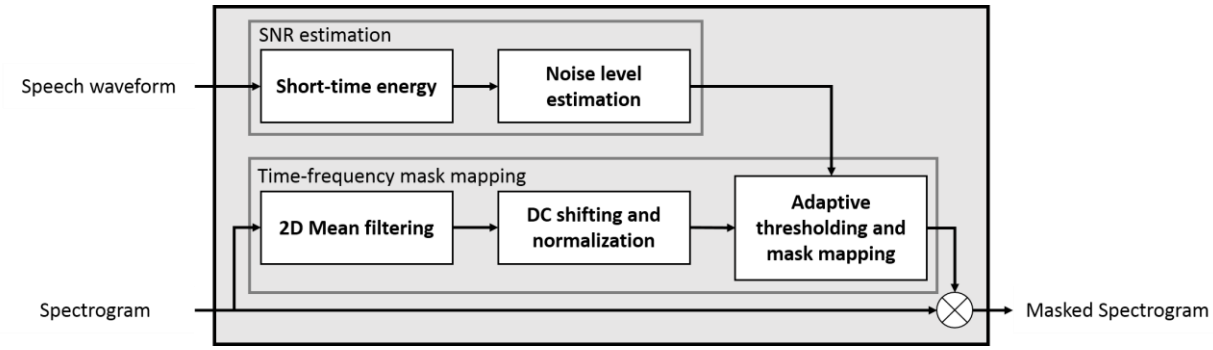


Figure 3-2: Adaptive time-frequency masking block diagram

3.1.1 SNR Estimation

The SNR estimation technique is applied to measure the noise power level within the speech signal [40]. The Global SNR (GSNR) for speech signal is defined as:

$$GSNR = \frac{\sigma_s^2}{\sigma_n^2} \qquad 3\text{-}1$$

where σ_s^2 is the power of signal and σ_n^2 is the power of noise. The standard SNR definition can be evaluated from GNSR, where the speech activity is detected from the speech signal using one of Voice Activity Detection (VAD) techniques. Therefore, equation (3-1) can be rewritten as:

$$SNR = 10\log \frac{\sum_{n=0}^{l-1} s^2[n] \cdot vad[n]}{\sum_{n=0}^{l-1} n^2[n].vad[n]} \qquad 3\text{-}2$$

where $s[n]$ is the speech samples, $n[n]$ is the noise samples, and $vad[n]$ is the detected voice activity within the speech waveform. The detection of silence and voice activity is a sensitive stage for correct noise level estimation. There are many approaches to find the voice activity information. One of these approaches is Energy-based algorithm [40], which is used, in the proposed method:

3.1.1.1 Short-time energy

Due to the quasi-stationary property of the speech signal, it is analyzed into short frames. The short time energy STE_i of the noisy speech frames can be defined as:

$$STE_i = \sum_{n=0}^{J-1} \left(x_i[n]\right)^2 = \sum_{n=0}^{J-1} \left(s_i[n] + n_i[n]\right)^2 \qquad 3\text{-}3$$

where n is the sample index, i is the frame index, J is the total number of the samples within the frame, $x_i[n]$ is noisy speech samples, $s_i[n]$ is speech samples and $n_i[n]$ is noise samples.

3.1.1.2 Noise level estimation

In the silence frames, the short time energy is considered as a noise; while the higher energy frames are considered is speech signals with additive noise in the voice activity frames.

Consequently, the power of the speech signal to the power of the noise can be rewritten in terms of short time energies as follows:

$$SNR = 20\log \frac{\sum_{i=1}^{T'}\sum_{n=0}^{J-1} s_i[n]}{\sum_{i=1}^{T'}\sum_{n=0}^{J-1} n_i[n]} \qquad 3\text{-}4$$

$$SNR = 20\log \frac{\sum_{i=1}^{T'}\sum_{n=0}^{J-1} x_i[n] - n_i[n]}{\sum_{i=1}^{T'}\sum_{n=0}^{J-1} n_i[n]} \qquad 3\text{-}5$$

where k is the total number of frames. The voice short energy varies over the time, while the additive noise energy is almost constant. Estimating the noise frames for the uttered word can be evaluated from the low short energy frames, while the higher short energy frames are considered as the voice activity frame with AWGN; as a result, the Estimated SNR (ESNR) is calculated by using the following relation:

$$ESNR = 20\log \frac{\sum_{i=1}^{T'} STE_i - T' \cdot \min\left(STE_i\right)}{T' \cdot \min\left(STE_i\right)} \qquad 3\text{-}6$$

The ESNR value is used later in adaptive thersholding and mask-mapping stage.

3.1.2 Time-frequency Mask Mapping

In the speech waveform, high frequencies are more sensitive to the additive noise since the most speech energies are concentrated in low frequencies. Therefore, the adaptive mask is constructed to concentrate the feature values in low frequencies rather than in high frequencies. The following algorithm is applied to extract the less corrupted information from the noisy speech waveform.

3.1.2.1 2D Average filtering

The aim of this stage is to construct a smooth shape of the spectrogram. It is implemented by convolving 11×11 uniform kernel functions with the time-frequency map. The normalized kernel function equations are given as:

$$I'(m,k) = \sum_{x=-5}^{5}\sum_{y=-5}^{5} 1 \times I(m+x, k+y) \qquad 3\text{-}7$$

$$I'_{norm.}(m,k) = \frac{I'(m,k)}{\sum_{x=-5}^{5}\sum_{y=-5}^{5} 1} \qquad 3\text{-}8$$

where *x* and *y* are the filter index and *m* and *k* are the spectrogram value indexes.

3.1.2.2 DC shifting and normalization

The scale of the constructed smoothed shape varies for each word. Therefore, it is DC shifted and zero floored by subtracting it from the smallest value, and then dividing it by the maximum number normalizes the total shape.

3.1.2.3 Adaptive thersholding and mask mapping

In this stage, the estimated SNR from Equation (3-6) is used to create an adaptive thresholding. The relation between the threshold and the estimated SNR is demonstrated experimentally by calculating the threshold values that produces the highest recognition rate at different estimated SNRs, then curve fitting is the applied on the obtained points as shown in Figure (3-3). Equation (3-9) is the derived formula from curve fitting process where a_{fit} and b_{fit} are constants with values 0.047 and 0.8, respectively. The threshold value is calculated to adapt noise level divergence. It is inversely proportional to the Estimated SNR value.

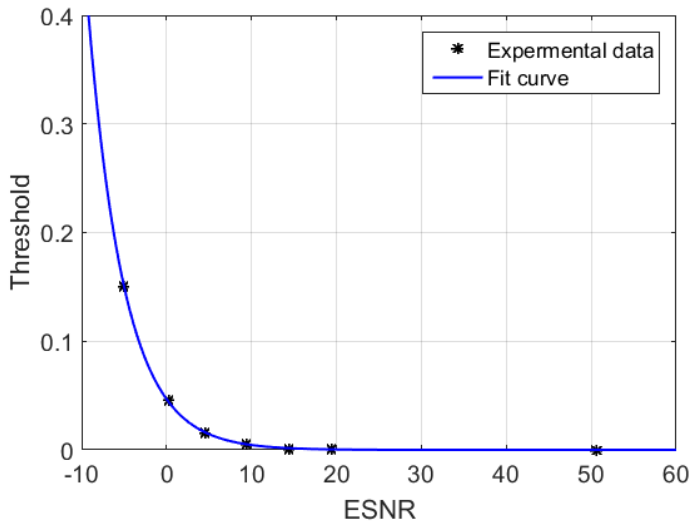


Figure 3-3: The threshold values at different Estimated Signal to Noise Ratios (ESNR)

$$Thershold = a_{fit}.b_{fit}^{ESNR} \qquad 3\text{-}9$$

The threshold value is applied on the DC shifted and normalized smoothed shape. If the smoothed shape values are greater than the threshold value, the less corrupted speech information is detected and the value of the mask map is equal to 1. Meanwhile, it is equal to 0.1 if the smoothed shape values are less than the threshold value, which means highly corrupted speech information with noise.

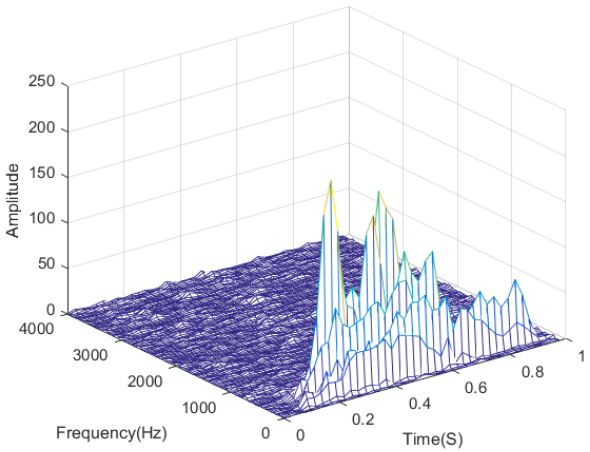
The time-frequency map is multiplied by constructed mask map, which results the low distorted speech information is multiplied by 1 and highly distorted speech information is multiplied by 0.1. Thus, the speech features are weighted in the less distorted values.

3.1.3 Mean Smoothing Filter

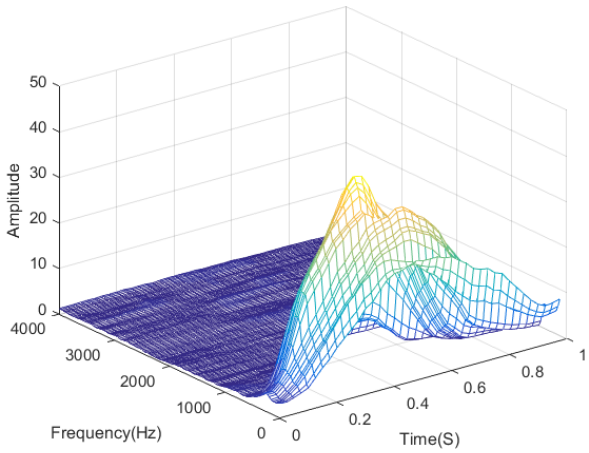
Since the power associated with AWGN varies differently from that associated with speech signal, the mean smoothing filter is applied on the adaptive masked time-frequency map to remove the instant variation on the power along the frames. The filter is implemented by calculating the average of three frequency frames. The filter kernel function is given by:

$$H(z) = \frac{1}{3}(1 + z^{-1} + z^{-2}) \qquad 3\text{-}10$$

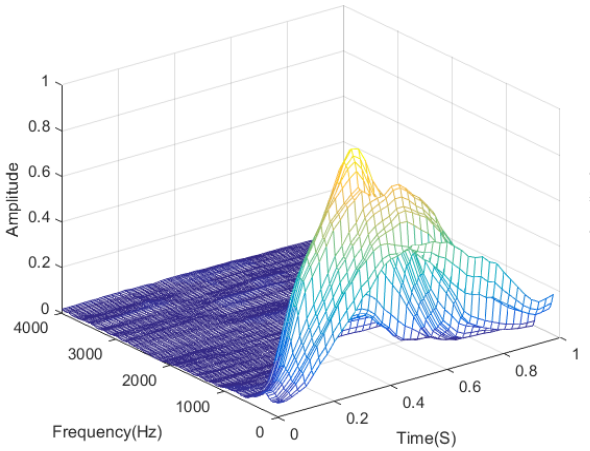
After mean smoothing filtering the triangular Mel-filter banks are applied then log DCT is applied to obtain MFCC features. The proposed method stages are illustrated graphically on the noisy word as shown in Figure (3-4)



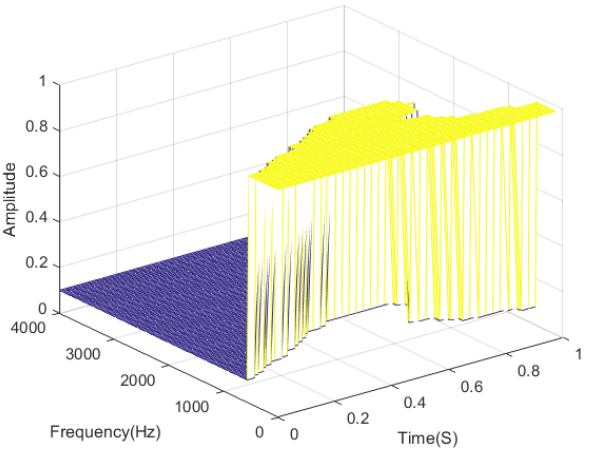
a) Noisy Spectrogram



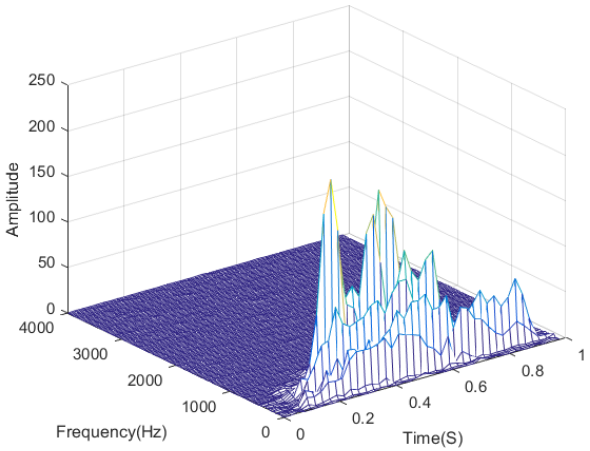
b) After 2D Average filtering



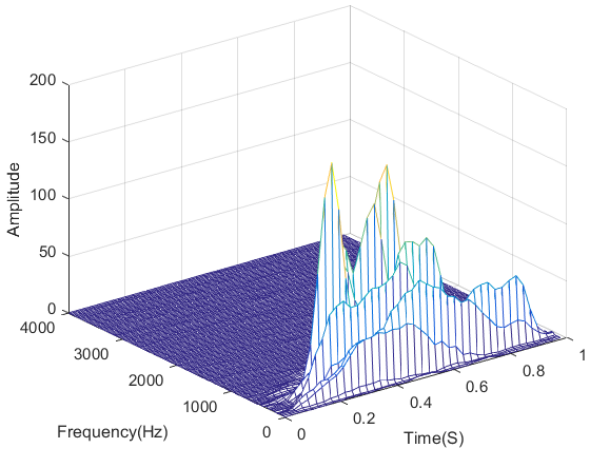
c) After DC shifting and normalization



b) Speech mask map



e) After multiplication by the mask



f) After Average smoothing

Figure 3-4: Spectrogram of the uttered word 'one' at SNR = 5 dB

3.2 Proposed Method 2: Modified PNCC Technique

The second proposed method was developed improve the PNCC system performance in the presence of Additive White Gaussian Noise (AWGN) and different types of environmental noise. The proposed feature extraction method was developed to obtain the acoustic features that can filter the non-stationary background noise without affecting the performance of undistorted speech signal. This system is implemented using 8 kHz speech dataset.

Figure (3-5) shows the block diagram of the proposed system. The pre-emphasize filter is applied on the input speech waveform. After that, the speech wave is divided into short overlapped frames in 25.6 ms frame duration with 10 ms overlap between frames and the each frame is multiplied by a hamming window. The PSD $|Y(k)|^2$ is obtained from each window by computing the magnitude-square of the FFT with 256 bit resolution.

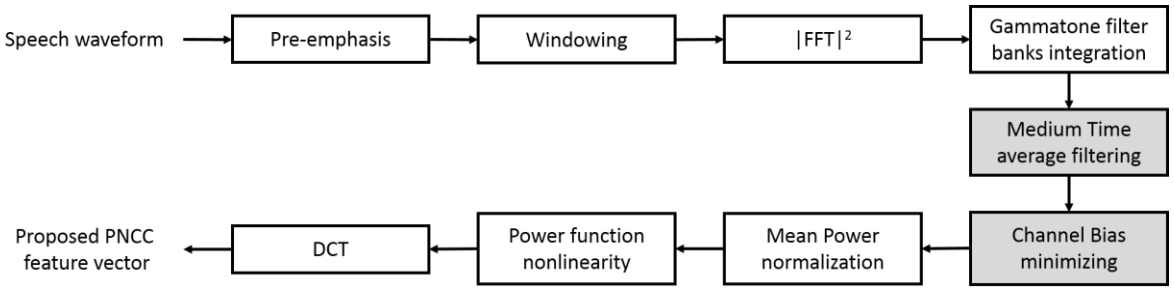


Figure 3-5: Block diagram of the second proposed system

Afterward, a set 25 of Gammatone filters were generated from the frequency response of the Gammatone kernel basis functions with center frequencies spanning from 100 Hz to 4 kHz as shown in Figure (3-6). Each filter bank is normalized and the value is substituted by zero if the filter bank is less than of 0.5 percent of its maximum value. Each Gammatone filter is multiplied by the PSD of each frame and the summation is calculated as in Equation (3-11).

$$P[m,l] = \sum_{k=1}^{M/2} |Y[m,k]|^2 G_l(k) \qquad 3\text{-}11$$

where l is the Gammatone filter index, m is the frame index, $G_l(k)$ is the function of each filter in frequency domain and M is spectrum resolution.

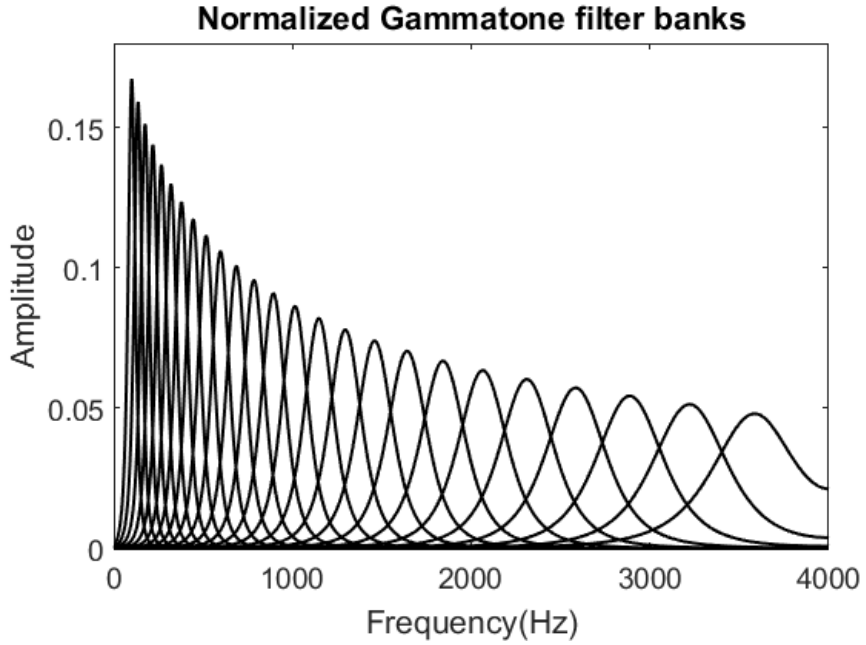


Figure 3-6: Normalized 25 Gammatone filter banks

3.2.1 Medium Time Average Filtering

Long analysis windows are usually used in the most speech recognition systems. It found that using long analysis windows improves the system performance [5, 34, 41, 42] because the power that produced from the background noise conditions changes slower than the power that produced from the speech. Moreover, it has found that the Gammatone channel envelopes can provide matching information that is used for enhancing noisy speech recognition accuracy as in many algorithms [43-45]. In this research, many filtering techniques types along frames were used. For instance, Gaussian filter, Laplacian filter, Median filter and Average filter. The highest performance is obtained in the case of Average filter. The medium-time power $Q[m,l]$ is an Average filtering technique by computing the running average power $P[m,l]$ for $2M+1$ consecutive frames as in the following equation.

$$\tilde{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} P[m',l] \qquad 3\text{-}12$$

where m represents the frame index and l is the Gammatone channel index. This stage is used differently from the PNCC system. The frames are low pass filtered and smoothing effect

between from filtering will remove the fast changes in amplitude that is associated with additive background noise. Using large average window will degrade the performance of will blur the speech data and will degrade the system performance. As well as, it will increase the computational time. Thus, the value of coefficient M is derived experimentally in the next section.

3.2.2 Channel Bias Minimizing

The most of the speech information is concentrated at low frequencies while the noise frequency distribution varies for each type of noise. The noisy speech spectrum is smoothed when it is multiplied by Gammatone filterbanks and it smoothed again along each Gammatone channel by the medium time average filtering. The large window smoothing effect will cause spreading the noise energies more than the speech energies along each Gammatone channel. The channel bias will be produced and it depends on the noise spectral distribution more than the speech spectral distribution. Therefore, the bias value varies from each Gammatone channel to another. In this stage, the channel bias effect is minimized by subtracting channel values from the minimum number within each channel multiplied by a factor d where $0 < \theta < 1$.

$$\breve{Q}[m,l] = \tilde{Q}[m,l] - \theta \times \min \tilde{Q}[l] \qquad 3\text{-}13$$

The value of d and M were derived experimentally in this research by choosing multiple values that can cause a high performance for the noisy speech. A non-Gaussian distributed environmental noise such as Exhibition noise. As shown in the Figure (3-7) in the case of Exhibition noise at SNR 5dB, the highest performance of the system at $M = 6$ with coefficient $\theta = 0.7$ and it is slightly higher than the performance at of $M = 5$ with coefficient $\theta = 0.6$. In Figure (3-8) at SNR -5dB, the performance of the system at $M = 6$ with coefficient $\theta = 0.7$ and it is still slightly higher than the performance at of $M = 5$ with coefficient $\theta = 0.6$. However, in the proposed system the value of $M = 5$ with coefficient $\theta = 0.6$ were used to reduce the processing time and to avoid losing the system performance due to blurring effect in the condition of uncorrupted speech.

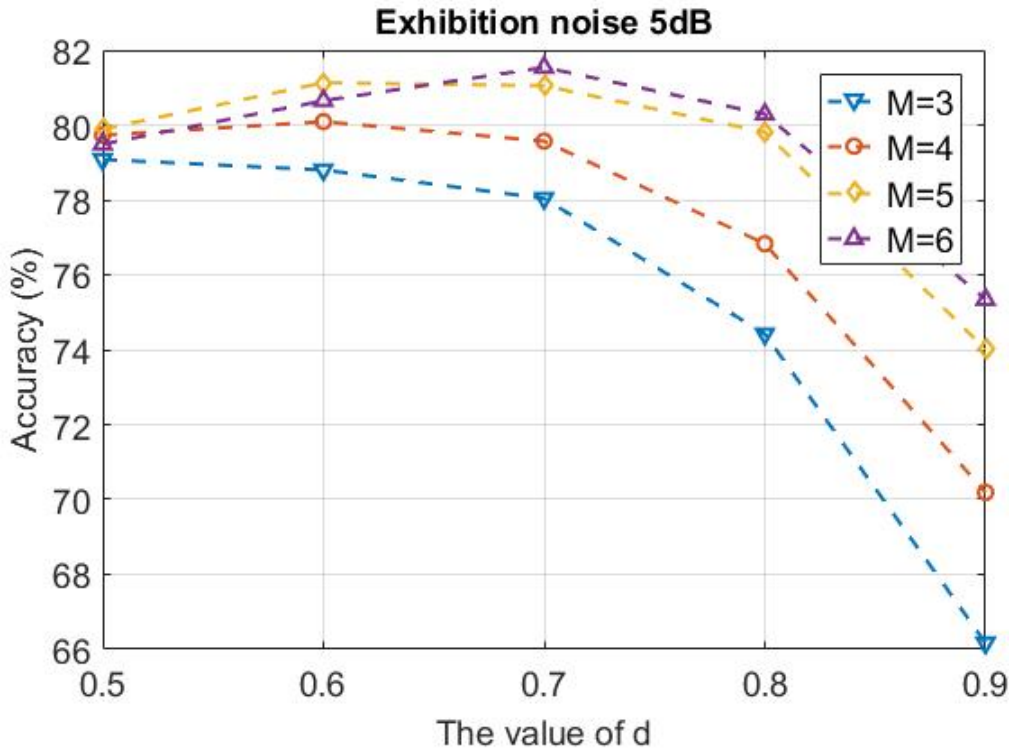


Figure 3-7: Recognition performance at different filter average width at 5dB

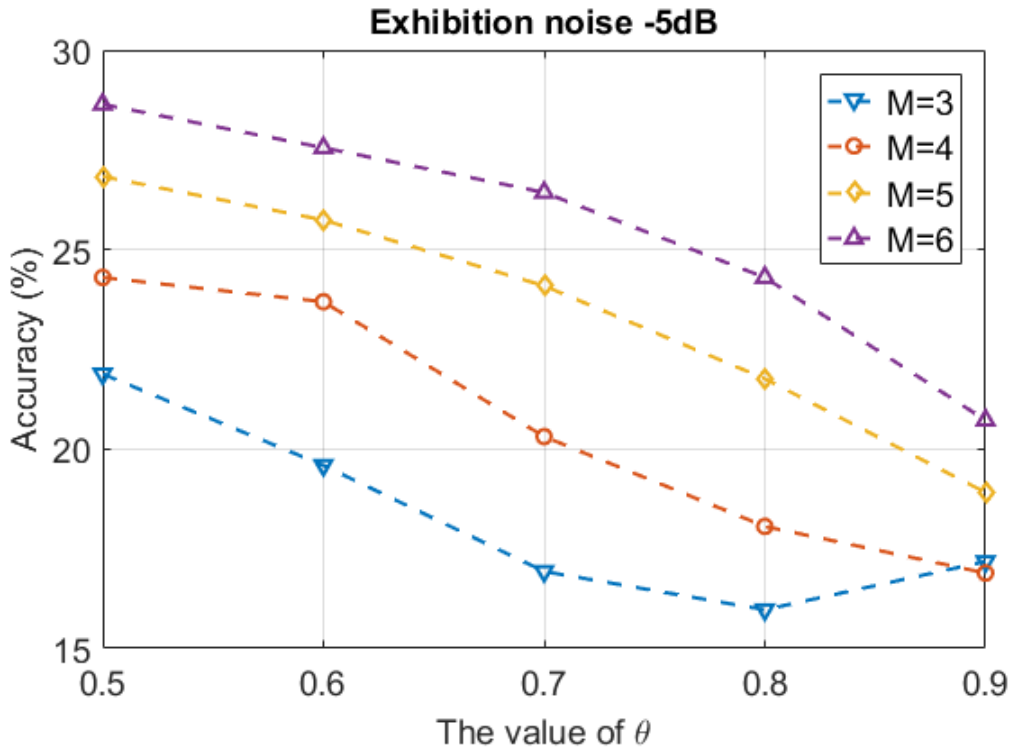


Figure 3-8: Recognition performance at different filter average width at -5dB

In the next stage, the mean power normalization is calculated by computing $\mu[m]$ from the equation:

$$\mu[m] = \lambda_\mu \mu[m-1] + \frac{(1-\lambda_\mu)}{L} \sum_{l=0}^{L-1} \breve{Q}[m,l] \qquad 3\text{-}14$$

where the forgetting factor λ_μ *is* 0.999. The normalized power is obtained from the estimated running power μ[m] as in the following equation:

$$U[m,l] = k \frac{\breve{Q}[m,l]}{\mu[m]} \qquad 3\text{-}15$$

Afterward, the power-law curve with an exponent of 1/15 is calculated. Lastly, the 13 proposed PNCC features are obtained by applying DCT.

$$V[m,l] = U[m,l]^{1/15} \qquad 3\text{-}16$$

CHAPTER FOUR

EXPERMENTAL WORKS AND RESULTS

4 Chapter FOUR: Experimental Works and Results

4.1 Database Description

The systems are implemented on excerpts of TIDIGITS Database [46] which was designed and collected by Texas Instruments (TI). It contains the of 11 digit sequences ("zero", "oh", "one", "two", "three", "four", "five", "six", "seven", "eight", and "nine"). The complete database consists of 326 speakers (111 men, 114 women, 50 boys, and 51 girls), each of them pronounces 22 isolated digits and 55 connected digits. The database is partitioned into two subsets. The first set is a training set and it consists of 55 men, 57 women, 25 boys, and 26 girls. The second set is a testing set and it consists of 56 men, 57 women 25 boys, and 25 girls. The words are recorded in 20 kHz.

In the excerpts of the database, the training set is 37 men and 57 women and the testing set is 56 men and 57 women. The isolated digits for these sets are utilized in both experiments to evaluate the performance of the two proposed method in comparison to the state of the art techniques. The records are downsampled to 8 kHz.

The database structure is single word. Therefore, the language model is only one word and in this case the recognition Accuracy is equal to the Word Recognition Rate (WRR).

4.2 Experimental Results of Proposed Method 1

In speech recognition system configuration, each word was framed by 25 ms overlapped Hamming windows, with 10 ms shifting between frames. The FFT resolution was 256 bit and the power spectrum of each frame was calculated. Then it is multiplied by 26 overlapped Mel-filter banks and 12 MFCC coefficients were calculated.

Models were trained with noise-free utterances, while tested with noise-free and noisy utterances. The Additive White Gaussian Noise (AWGN) was generated in a fixed sequence at different SNRs from -5 dB to 20 dB with step size 5dB. The experiment was implemented using standard MFCC, RASTA-PLP [47] and the proposed method MFCC without pre-emphasizing. As shown in Figure (4-1), the log energy, Δ and $\Delta\Delta$ features were calculated for the three systems; the total number of extracted features was 39 in each case. The HMM is used in classification stage. To calculate the number of states in each word, the Sphinx CMU lexical dictionary [48] that contains the phones of each word was used. The number of Markov states was 3 per each phone. The

number of HMM iterations was 10 and WRR was calculated for each iteration. Then, the average recognition rate of the last 5 WRR was calculated.

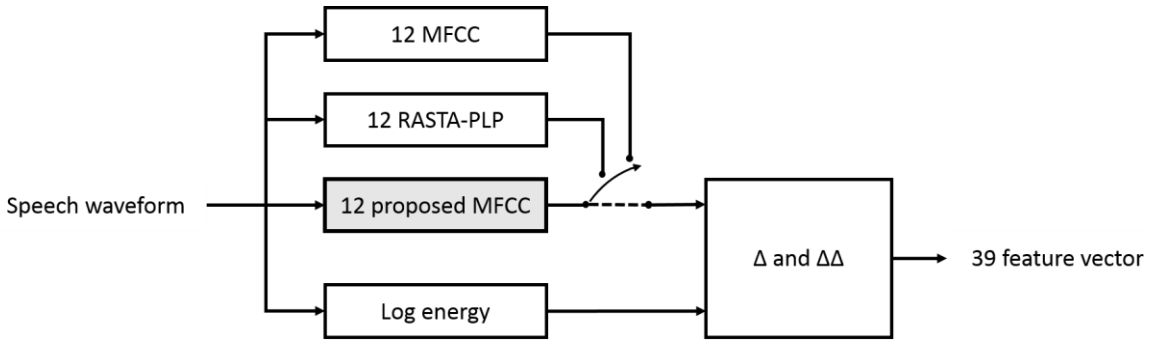


Figure 4-1: Feature extraction stage for the first experiment

The approached MFCC method is evaluated by comparing its recognition accuracy with MFCC and RASTA-PLP system. The performance of MFCC, RASTA-PLP, and proposed method is presented in clean data and in different SNRs from -5 dB to 20 dB with step size 5dB. The recognition accuracy is presented graphically as in Figure (4-2) and numerically as in Table (4-1).

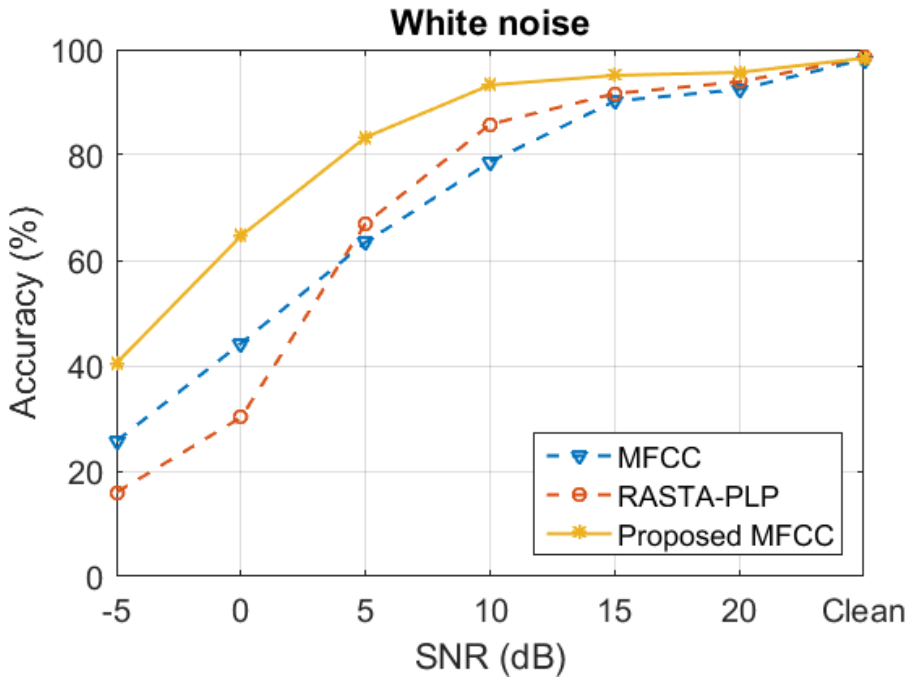


Figure 4-2: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for AWGN in the first proposed method

In Table (4-1), the proposed method outperform other methods in terms of recognition rate at all SNRs. For undistorted data, the recognition performance is almost constant. For low distorted data, the recognition accuracy is improved by 3.24% and 1.78% in case of 20 dB, 4.87% and 3.44% in the case of 15 dB and 14.57% and 7.45% in the case of 10 dB compared to MFCC and RASTA-PLP, respectively. For the high-distorted data, the recognition accuracy is improved by 19.92% and 16.37% in case of 5 dB, 20.37% and 34.45% in the case of 0 dB and 14.79% and 24.52% in the case of -5 dB compared to MFCC and RASTA-PLP, respectively. The obtained processing time for the same uttered word "one" is 0.012s, 0.053s and 0.035s for MFCC, RASTA-PLP and proposed MFCC, respectively.

Table 4-1: Percentage Word Recognition Rate (WRR) for AWGN in the first proposed method

SNR	**White noise**		
	MFCC	**RASTA-PLP**	**Proposed MFCC**
-5 dB	25.70	15.97	40.92
0 dB	44.27	30.19	64.64
5 dB	63.44	66.98	83.36
10 dB	78.71	85.83	93.28
15 dB	90.23	91.65	95.09
20 dB	92.43	93.89	95.67
Clean	98.18	98.48	98.36

4.3 Experimental Results of Proposed Method 2

The experiment was implemented using MFCC, RASTA-PLP [47], GFCC, PNCC, and the proposed modified PNCC systems. In the configuration of the speech recognition system, each word was framed by 25.6 ms overlapped Hamming windows, with 10 ms shifting between frames. The FFT resolution was 256 bit and the power spectrum of each frame was calculated. For the MFCC system, 26 overlapped Mel-filter banks and 13 MFCC coefficients were calculated. In the case of GFCC, PNCC and proposed PNCC systems, 25 Gammatone filter banks were generated from Gammatone frequencies in the range of 100 Hz to 4 kHz. The Gammatone filters are normalized and their values are equal to zero if it is less than 5 percent of the maximum

value for each filter. The Equation (2-36) is used in integrating the PSD with the Gammatone filter banks in PNCC. While the Equation (3-11) is used in GFCC and proposed PNCC systems.

Models were trained with noise-free utterances, while tested with noise-free and noisy utterances. The eight different types of noise were used one of them is Additive White Gaussian Noise (AWGN) and the rest are environmental noise such as "Airport", "Babble", "Car", "Exhibition", "Restaurant", "Street" and "Subway". These noises are described in Table (4-2). All of these noises were added to testing datasets with different SNRs from -5 dB to 20 dB with step size 5dB.

Table 4-2: Noise description

Noise type	Source of noise
AWGN	Wideband noise generated from many natural sources
Babble	Mixture of numerous human voices
Airport	Ambience from an airport lobby
Restaurant	Environment of a typical restaurant
Exhibition	Ambience from an exhibition hall
Street	Ambience outdoors on a city street
Car	Noise inside a moving car
Subway	Noise inside a moving subway train

As shown in Figure (4-3), the Cepstral Mean Normalization (CMN) is applied. Next, Δ and ΔΔ features were calculated for the five systems; the total number of extracted features was 39 in each case.

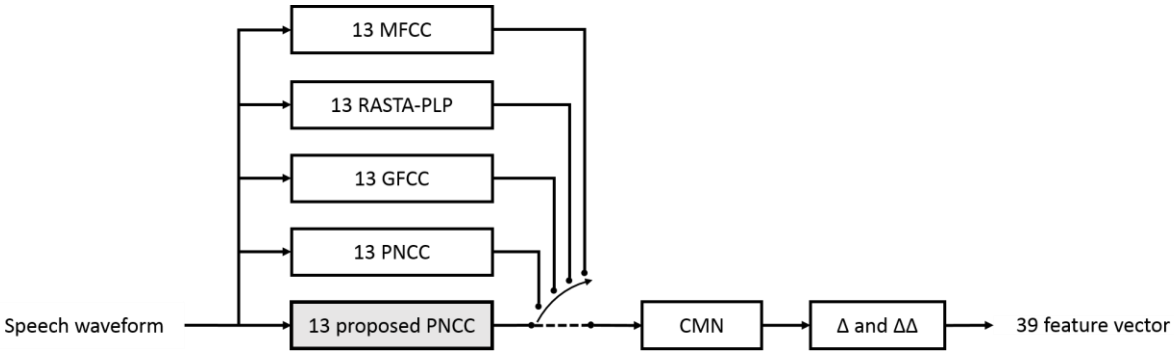


Figure 4-3: Feature extraction stage of the second experiment

Similarly, as in the first proposed method, The HMM is also used in classification stage. To calculate the number of states in each word, the Sphinx CMU lexical dictionary [48] was used. The number of Markov states was 3 per each phone. The number of Markov chains was 3 per phoneme. The number of HMM iterations was 10 and WRR were calculated for each iteration and the maximum recognition rate was selected.

The proposed method is evaluated by comparing its recognition accuracy with MFCC and RASTA-PLP, GFCC and PNCC methods. The performance of MFCC, RASTA-PLP, GFCC, PNCC and the proposed method is presented in clean data and in different SNRs from -5 dB to 20 dB with step size 5dB. The recognition accuracy is presented graphically and numerically for different types of noise. As observed in the all the recognition rate figures and tables at different types of noise, the average recognition rate is dramatically improved in low Signal to Noise Ratios, whereas it is higher than or almost unchanged at high Signal to Noise Ratios.

The percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for AWGN is illustrated graphically in Figure (4-4). The recognition rate is extremely improved at low SNRs. At SNR-5 dB, and the recognition performance is improved by 29.97%, 26.01%, 24.53% and 11.73% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is enhanced by 47.15%, 44.57%, 29.16% and 7.04% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance still the highest by 46.14%, 44.73%, 22.12% and 1.65% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 22.65%, 23.13% and

10.98% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. At SNR 15 dB, the recognition performance is improved by 11.58%, 8.2% and 4.14% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. At SNR 20 dB, the recognition performance is improved by 3.42%, 1.69% and 1.89% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. While the performance is slightly less than PNCC method by 0.52%, 1.05% and 1.41% at 10 dB, 15 dB, and 20 dB, respectively.

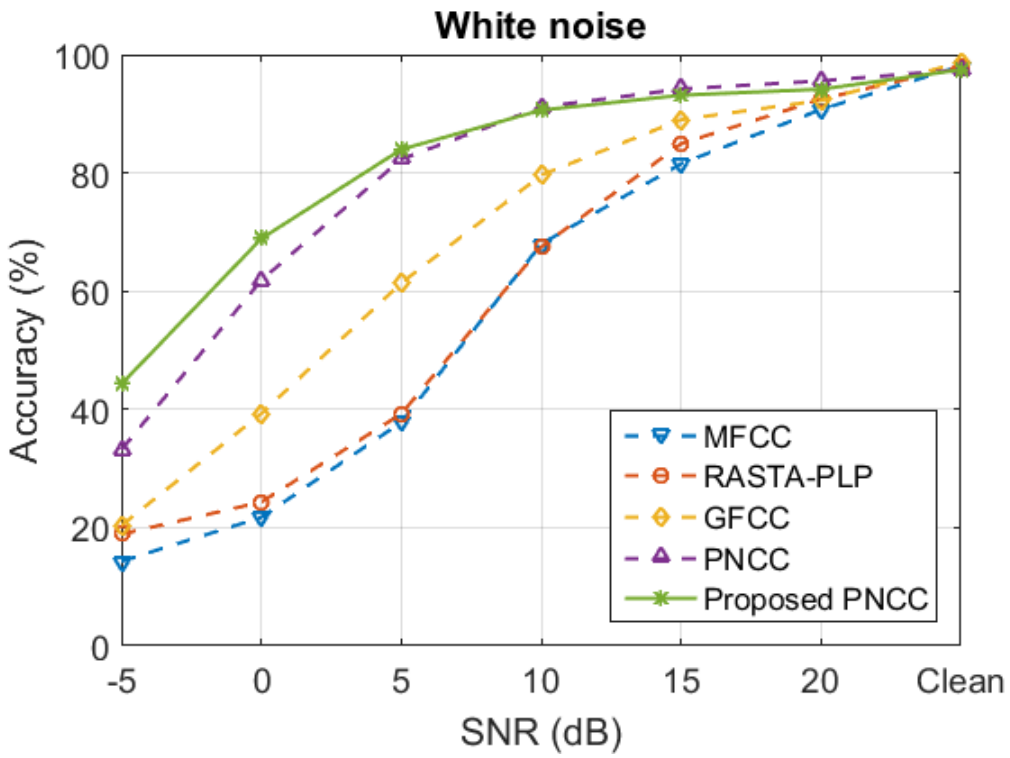


Figure 4-4: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for AWGN in the second proposed method

Table 4-3: Percentage Word Recognition Rate (WRR) for AWGN in the second proposed method

SNR	White noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	14.32	18.91	20.39	33.19	44.29
0 dB	21.76	24.34	39.30	61.87	68.91
5 dB	37.85	39.26	61.38	82.34	83.99
10 dB	67.98	67.50	79.65	91.15	90.63
15 dB	81.58	84.96	89.02	94.21	93.16
20 dB	90.75	92.48	92.28	95.58	94.17
Clean	98.11	98.03	98.75	97.47	97.47

The performance of the implemented systems at different Signal to Noise Ratios (SNR) in the Airport noise condition are shown in Figure (4-5). At SNR -5 dB, the recognition performance is improved by 24.82%, 24.22%, 27.35% and 6.96% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is increased by 39.58%, 39.83%, 26.79% and 6.56% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance is improved by 31.61%, 36.04%, 9.37% and 1.69% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is enhanced by 12.3%, 22.76% and 3.29% in comparison to MFCC, RASTA-PLP, and GFCC methods respectively while it is less than PNCC method by 0.25%. At SNR 15 dB, the recognition performance is improved by 0.16%, 3.62% and 0.24% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively while it is less than PNCC system by 1.36%. At SNR 20 dB, the recognition performance is slightly degraded by 0.89%, 0.81%, 0.2% and 0.64% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively.

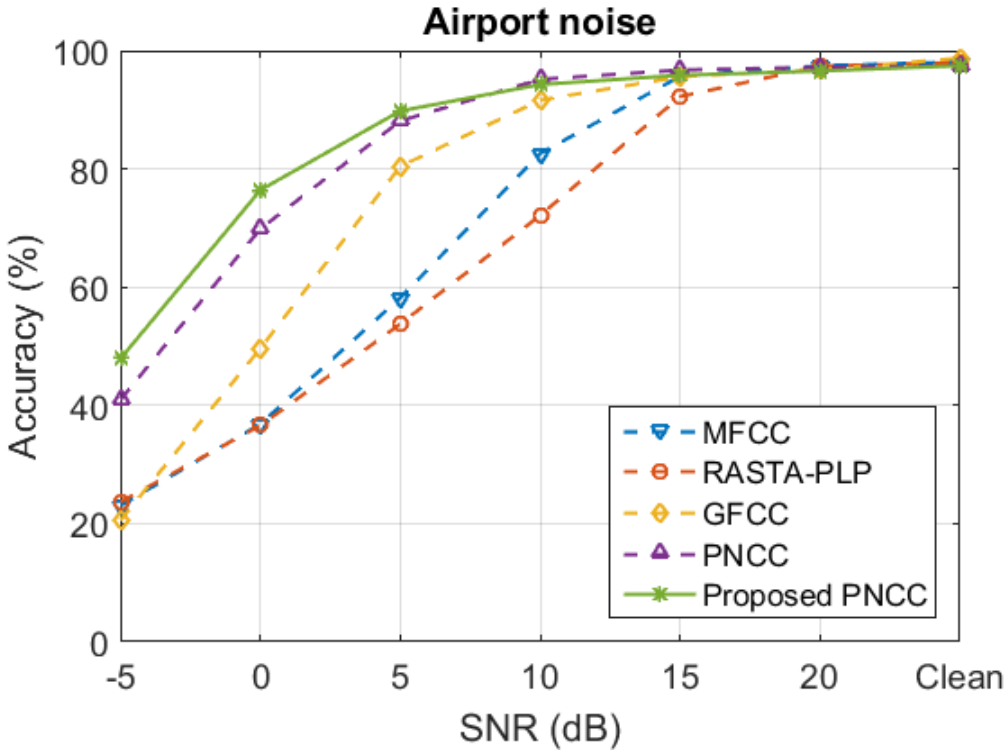


Figure 4-5: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for

Table 4-4: Percentage Word Recognition Rate (WRR) for Airport noise

SNR	Airport noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	23.09	23.69	20.56	40.95	47.91
0 dB	36.85	36.60	49.64	69.87	76.43
5 dB	58.21	53.78	80.45	88.13	89.82
10 dB	82.62	72.16	91.63	95.17	94.29
15 dB	95.70	92.24	95.62	96.74	95.86
20 dB	97.47	97.39	96.78	97.22	96.58
Clean	98.11	98.03	98.75	97.47	97.47

In the Babble noise case, the performance of the implemented systems at different Signal to Noise Ratios (SNR) are illustrated in Figure (4-6). At SNR -5 dB, the recognition performance is improved by 5.87%, 11.58%, 12.75% and 1.48% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is improved by 25.5%, 31.53%, 26.47% and 4.3% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance is improved by 33.63%, 39.3%, 15.69% and 3.74% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 18.27%, 29.21%, 5.27% and 1.49% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 15 dB, the recognition performance is improved by 2.54%, 10.58% and 1.01% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively while it is less than PNCC method by 0.2%. In the case of SNR 20 dB, the recognition performance is improved by 0.76% in comparison to RASTA-PLP method and 0.04% in comparison to GFCC method while is slightly degraded by 0.24% in comparison to MFCC method and by 0.04% in comparison to PNCC method.

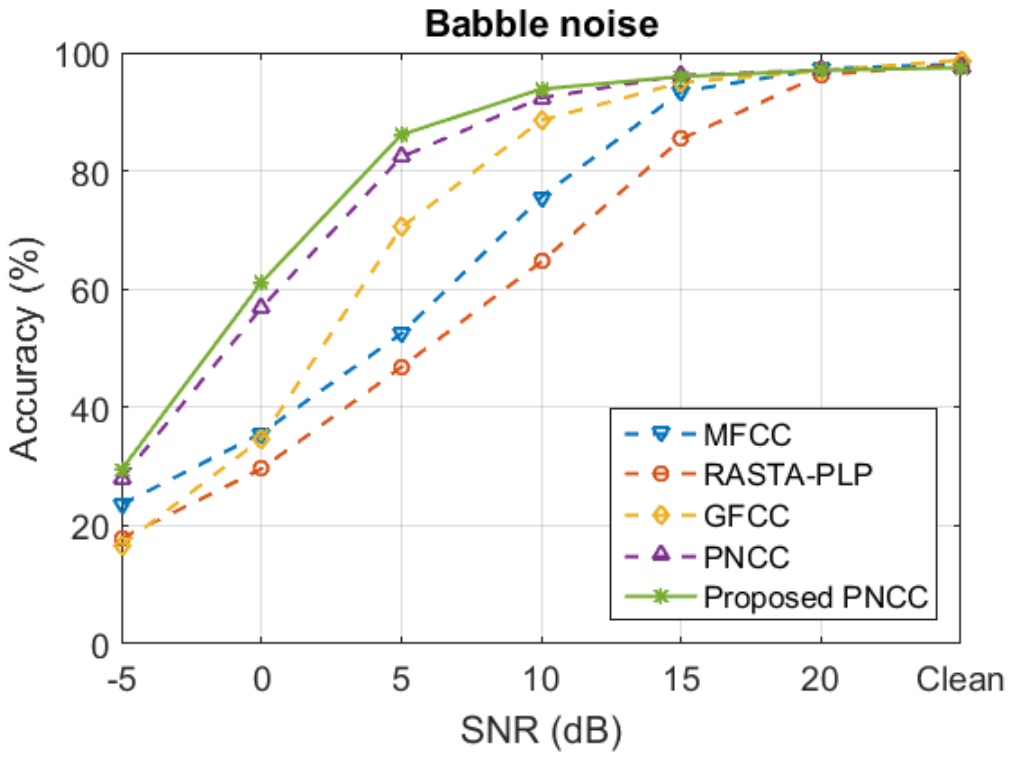


Figure 4-6: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Babble noise

Table 4-5: Percentage Word Recognition Rate (WRR) for Babble noise

SNR	Babble noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	23.57	17.86	16.69	27.96	29.44
0 dB	35.64	29.61	34.67	56.84	61.14
5 dB	52.49	46.82	70.43	82.38	86.12
10 dB	75.58	64.64	88.58	92.36	93.85
15 dB	93.44	85.40	94.97	96.18	95.98
20 dB	97.26	96.26	96.98	97.06	97.02
Clean	98.11	98.03	98.75	97.47	97.47

The performance of the implemented systems in the Car noise condition are shown in Figure (4-7). The performance of the proposed system is better than other systems at different Signal to Noise Ratios (SNR). In the case of SNR -5 dB, the recognition performance is improved by 33.75%, 33.79%, 36.97% and 19.51% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is improved by 48.19%, 46.98%, 40.02% and 12.83% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. In the case of SNR 5 dB, the recognition performance is improved by 37.98%, 42.68%, 18.06% and 3.99% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 13.36%, 27.07% and 6.4% in comparison to MFCC, RASTA-PLP and GFCC methods respectively while it is less than PNCC method by 0.08% .At SNR 15 dB, the recognition performance is improved by 0.2%, 4.22% and 1.36% in comparison to MFCC, RASTA-PLP, and GFCC methods, respectively while it is less than PNCC method by 0.49%. At SNR 20 dB, the recognition performance is slightly degraded by 0.52%, 0.64% and 0.68% in comparison to MFCC, RASTA-PLP and PNCC methods, respectively while it is improved by 0.36% in comparison to GFCC method.

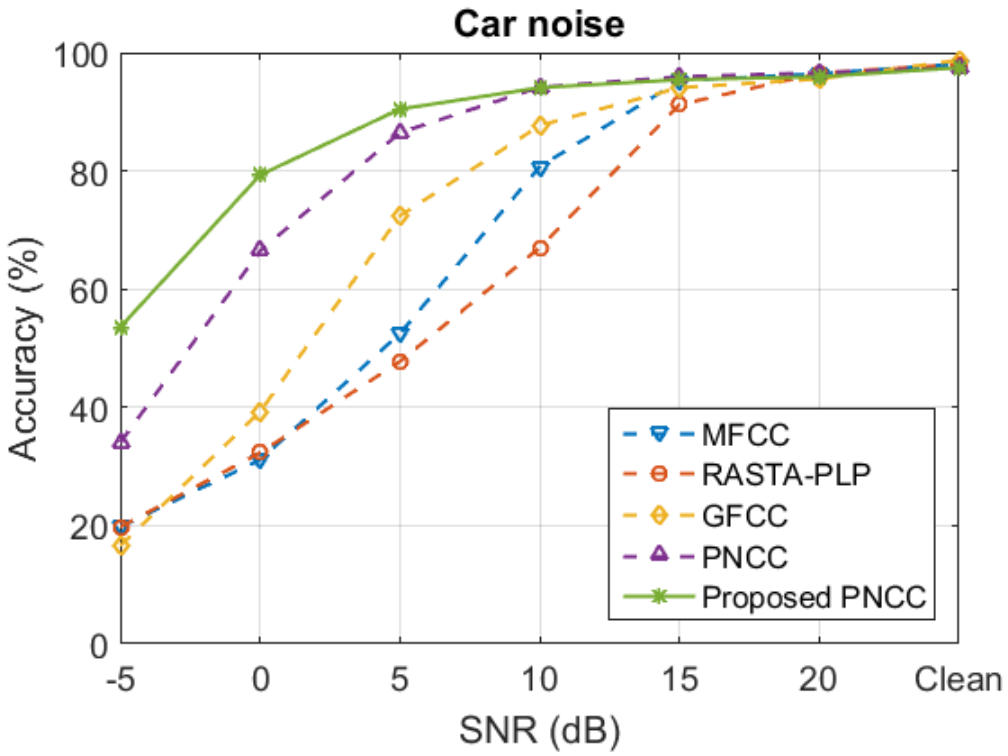


Figure 4-7: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Car noise

Table 4-6: Percentage Word Recognition Rate (WRR) for Car noise

SNR	Car noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	19.83	19.79	16.61	34.07	53.58
0 dB	31.13	32.34	39.30	66.49	79.32
5 dB	52.49	47.79	72.41	86.48	90.47
10 dB	80.73	67.02	87.69	94.17	94.09
15 dB	95.25	91.23	94.09	95.94	95.45
20 dB	96.46	96.58	95.58	96.62	95.94
Clean	98.11	98.03	98.75	97.47	97.47

In Figure (4-8), the performance of the implemented systems in the presence of Exhibition noise are illustrated. The recognition rate is enhancement is obtained at different SNRs. At SNR -5 dB, the recognition performance is improved by 11.22%, 10.53%, 9.05% and 0.84% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is improved by 35.39%, 28.44%, 24.17% and 6.99% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance is improved by 44.85%, 36.08%, 24.85% and 6.03% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 29.57%, 22.33%, 13% and 2.82% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 15 dB, the recognition performance is improved by 8.09%, 4.67% and 3.78% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. At SNR 20 dB, the recognition performance is improved by 1.01%, 0.45% and 1.17% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. While the performance is slightly less than PNCC method by 0.12% at SNR 15 dB and 0.68% at SNR 20 dB.

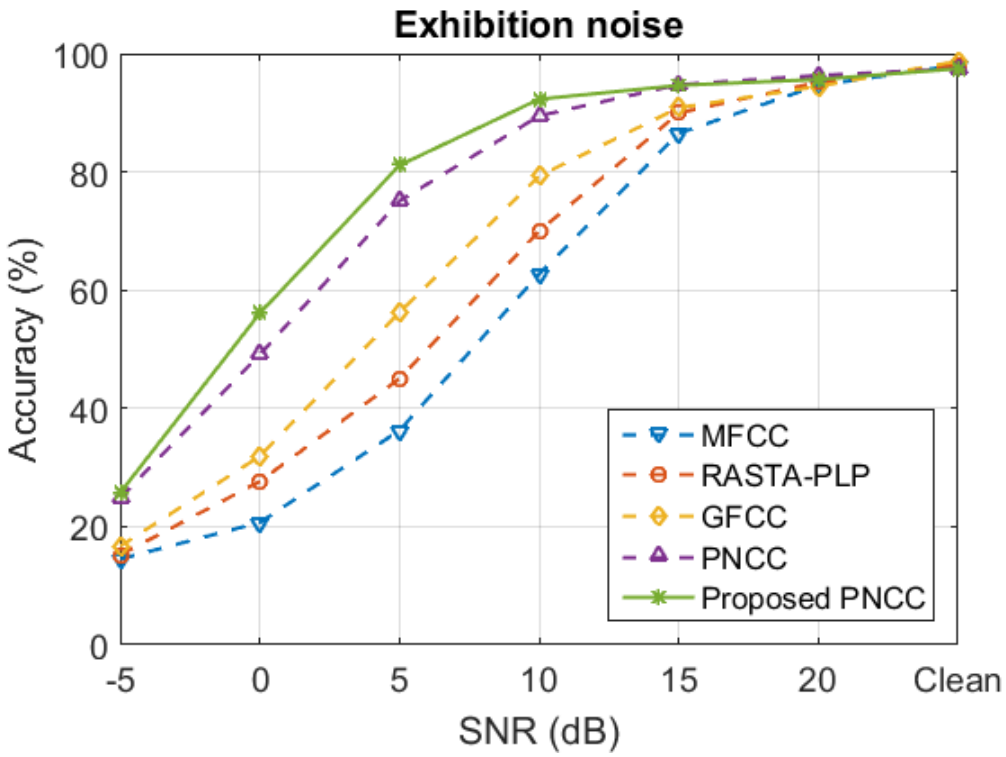


Figure 4-8: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Exhibition noise

Table 4-7: Percentage Word Recognition Rate (WRR) for Exhibition noise

SNR	Exhibition noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	14.52	15.21	16.69	24.90	25.74
0 dB	20.72	27.67	31.94	49.12	56.11
5 dB	36.28	45.05	56.28	75.10	81.13
10 dB	62.75	69.99	79.32	89.50	92.32
15 dB	86.56	89.98	90.87	94.77	94.65
20 dB	94.65	95.21	94.49	96.34	95.66
Clean	98.11	98.03	98.75	97.47	97.47

In the Restaurant noise condition, the percentage recognition rate is shown in Figure (4-9). At SNR - 5 dB, the recognition performance is improved by 6.36%, 9.61%, 11.18% and 0.52% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. In the case of SNR 0 dB, the recognition performance is improved by 26.07%, 29.17%, 23.42% and 4.03% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance is improved by 31.94%, 35.12%, 13.92% and 3.02% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 19.43%, 24.82%, 5.23% and 1.45% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 15 dB, the recognition performance is improved by 2.01%, 4.58% and 0.72% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively. On the other hand, it decreased by 0.4% in comparison to PNCC method. At SNR 20 dB, the recognition performance is also decreased by 0.52%, 0.24% and 0.2% in comparison to MFCC, RASTA-PLP and PNCC methods, respectively while it is improved by 0.04% in comparison to GFCC method.

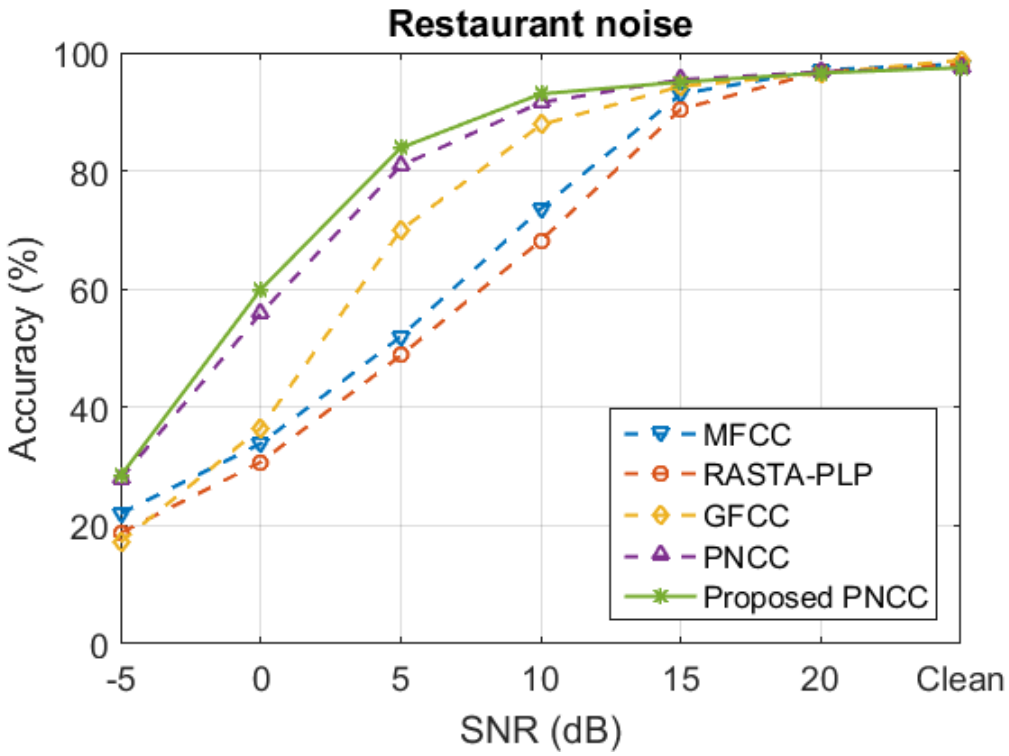


Figure 4-9: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Restaurant noise

Table 4-8: Percentage Word Recognition Rate (WRR) for Restaurant noise

SNR	Restaurant noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	22.04	18.79	17.22	27.88	28.40
0 dB	33.87	30.77	36.52	55.91	59.94
5 dB	51.97	48.79	69.99	80.89	83.91
10 dB	73.61	68.22	87.81	91.59	93.04
15 dB	93.04	90.47	94.33	95.45	95.05
20 dB	97.10	96.82	96.54	96.78	96.58
Clean	98.11	98.03	98.75	97.47	97.47

In the Street noise condition, the percentage recognition rate is presented in Figure (4-10). At SNR -5 dB, the recognition performance is improved by 32.26%, 37.21%, 38.66% and 15.53% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is improved by 38.82%, 45.05%, 30.33% and 8.61% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 5 dB, the recognition performance is improved by 27.72%, 37.37%, 8.65% and 1.29% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 11.02%, 26.51% and 1.97% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively while the performance is reduced in the case of PNCC method by 0.68%. At SNR 15 dB, the recognition performance is improved by 1.53% in comparison to MFCC method and 7.97% in comparison to RASTA-PLP method. On contrary, the performance is decreased by 0.12% in comparison to GFCC and by 0.64% in comparison to PNCC. At SNR 20 dB, the recognition performance is decreased by 0.16%, 0.2% and 0.36% in comparison to MFCC, GFCC and PNCC methods, respectively while the performance is enhanced in the case of RASTA-PLP method by 0.4%.

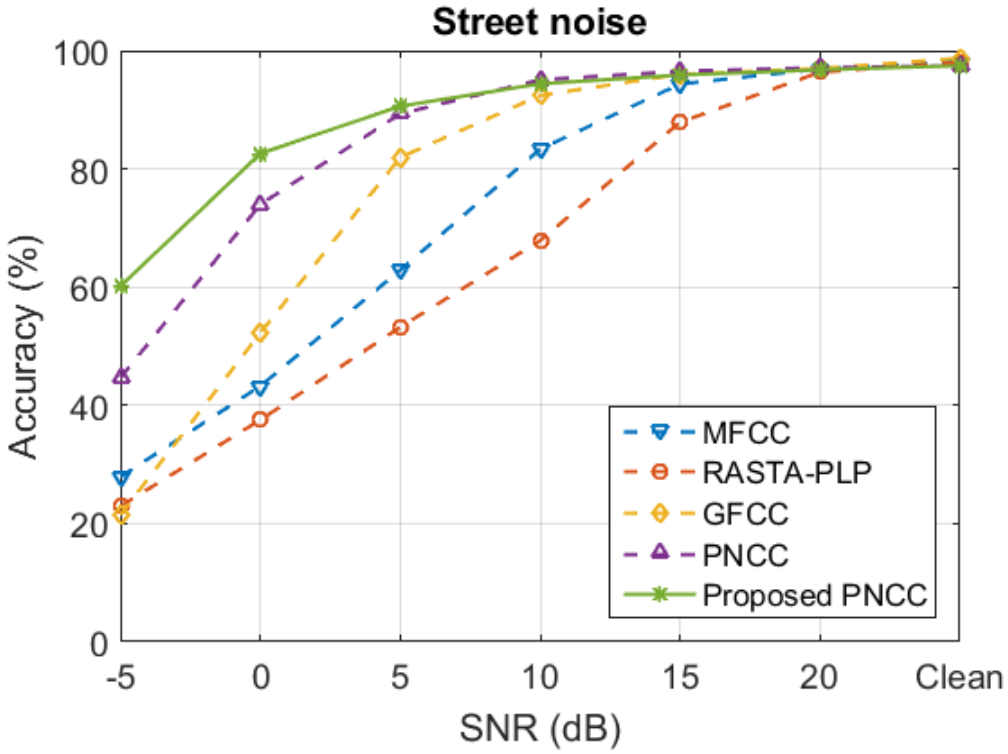


Figure 4-10: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Street noise

Table 4-9: Percentage Word Recognition Rate (WRR) for Street noise

SNR	Street noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	27.92	22.97	21.52	44.65	60.18
0 dB	43.28	37.57	52.29	74.01	82.62
5 dB	62.91	53.26	81.98	89.34	90.63
10 dB	83.39	67.90	92.44	95.09	94.41
15 dB	94.37	87.93	96.02	96.54	95.90
20 dB	96.94	96.38	96.98	97.14	96.78
Clean	98.11	98.03	98.75	97.47	97.47

Finally, in the case of Subway noise is illustrated in Figure (4-11), the recognition rate is enormously improved at low at SNR -5 dB, it improved by 26.02%, 25.5%, 22.69% and 11.66% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 0 dB, the recognition performance is improved by 49.75%, 45.25%, 32.34% and 14.4% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At 5 SNR dB, the recognition performance is improved by 55.72%, 47.87%, 22.57% and 8.16% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 10 dB, the recognition performance is improved by 35.04%, 33.95%, 7.81% and 1.09% in comparison to MFCC, RASTA-PLP, GFCC and PNCC methods, respectively. At SNR 15 dB, the recognition performance is improved by 8.37%, 9.41% and 2.98% in comparison to MFCC, RASTA-PLP and GFCC methods, respectively while the performance is reduced in the case of PNCC method by 0.16%. In the case of SNR 20 dB, the recognition performance is decreased by 0.81%, 0.45% and 0.93% in comparison to MFCC, RASTA-PLP and PNCC methods, respectively while the performance is enhanced in the case of GFCC method by 1.57%.

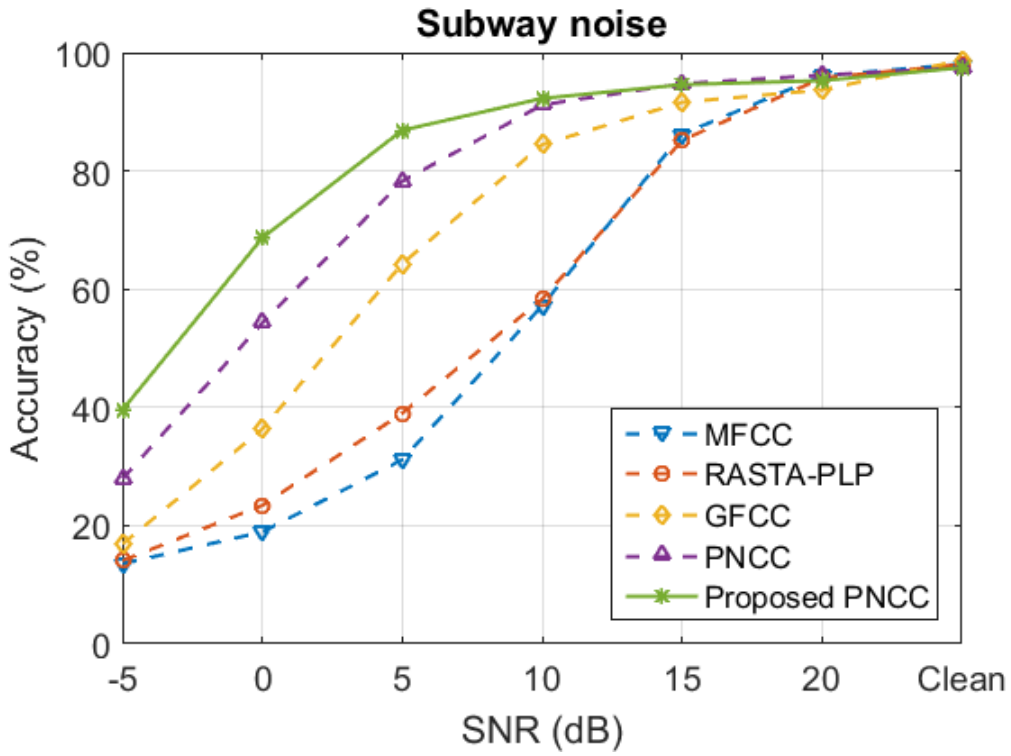


Figure 4-11: Percentage Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR) for Subway noise

Table 4-10: Percentage Word Recognition Rate (WRR) for Subway noise

SNR	Subway noise				
	MFCC	**RASTA-PLP**	**GFCC**	**PNCC**	**Proposed PNCC**
-5 dB	13.56	14.08	16.89	27.92	39.58
0 dB	18.95	23.45	36.36	54.30	68.70
5 dB	31.17	39.02	64.32	78.28	86.89
10 dB	57.24	58.33	84.47	91.19	92.28
15 dB	86.24	85.20	91.63	94.77	94.61
20 dB	96.10	95.74	93.72	96.22	95.29
Clean	98.11	98.03	98.75	97.47	97.47

The bar charts in Figure (4-12) demonstrate the percentage average improvement rate of the proposed method compared to the state-of-art methods at SNR -5 dB. As illustrated in the figure, the proposed method outcomes the other methods in the presence of white and environmental noise at very low SNR. The overall improvement in percentage recognition rate for all the types of noise is obtained in the case of MFCC, RASTA-PLP and GFCC methods, alternately. Followed by PNCC method. The highest recognition rate compared to RASTA-PLP and GFCC methods are determined in the Street noise condition. The proposed method is better than RASTA-PLP method by 37.21% while it is better than GFCC method by 38.66%. Differently, the highest recognition rate compared to MFCC and PNCC methods is determined in the case of Car noise. The proposed method is improved by 33.75% in the case of MFCC method while it enhanced by 19.51% more than PNCC method.

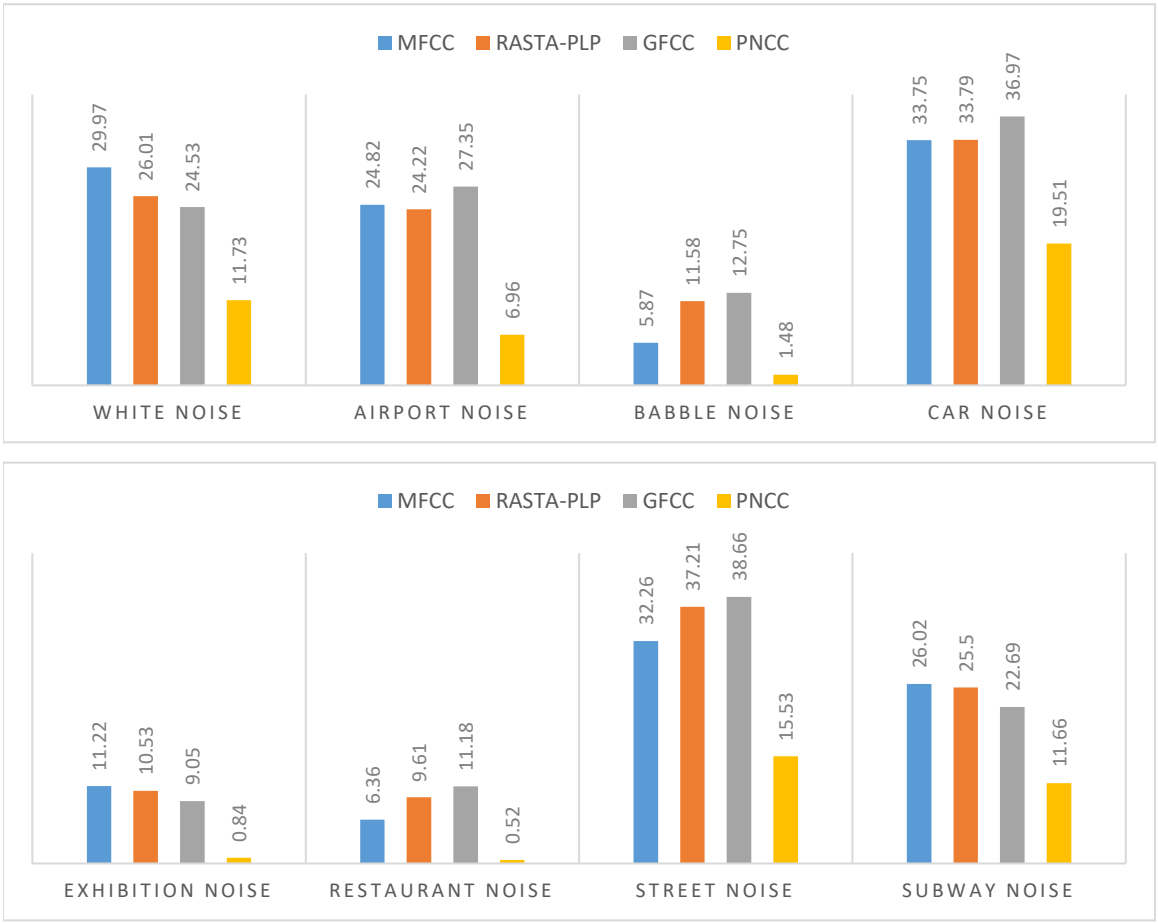


Figure 4-12: Percentage improvement rate for all types of noise at SNR -5dB

The percentage improvement rate of the proposed method compared to the other methods at SNR 0 dB is shown in bar charts in Figure (4-13). As shown in the figure, the proposed method is better than the other methods in the presence of white and environmental noise. In the Babble noise condition, the highest enhancement in percentage recognition rate is obtained in the case of RASTA-PLP method followed by GFCC and MFCC methods and lastly PNCC method. In the rest noises, the percentage recognition rate outperforms other methods in the case of MFCC and RASTA-PLP methods, alternately. Followed by GFCC method and finally PNCC method. The highest recognition rate compared to MFCC and PNCC methods are obtained in the Subway noise condition. The proposed method is better than MFCC method by 49.75% while it is better than PNCC method by 14.4%. Otherwise, the highest recognition rate compared to RASTA-PLP and GFCC methods is determined in the case of Car noise. The proposed method is improved by 46.98% in the case of RASTA-PLP method, whilst it enhanced by 40.02% in the case of GFCC method.

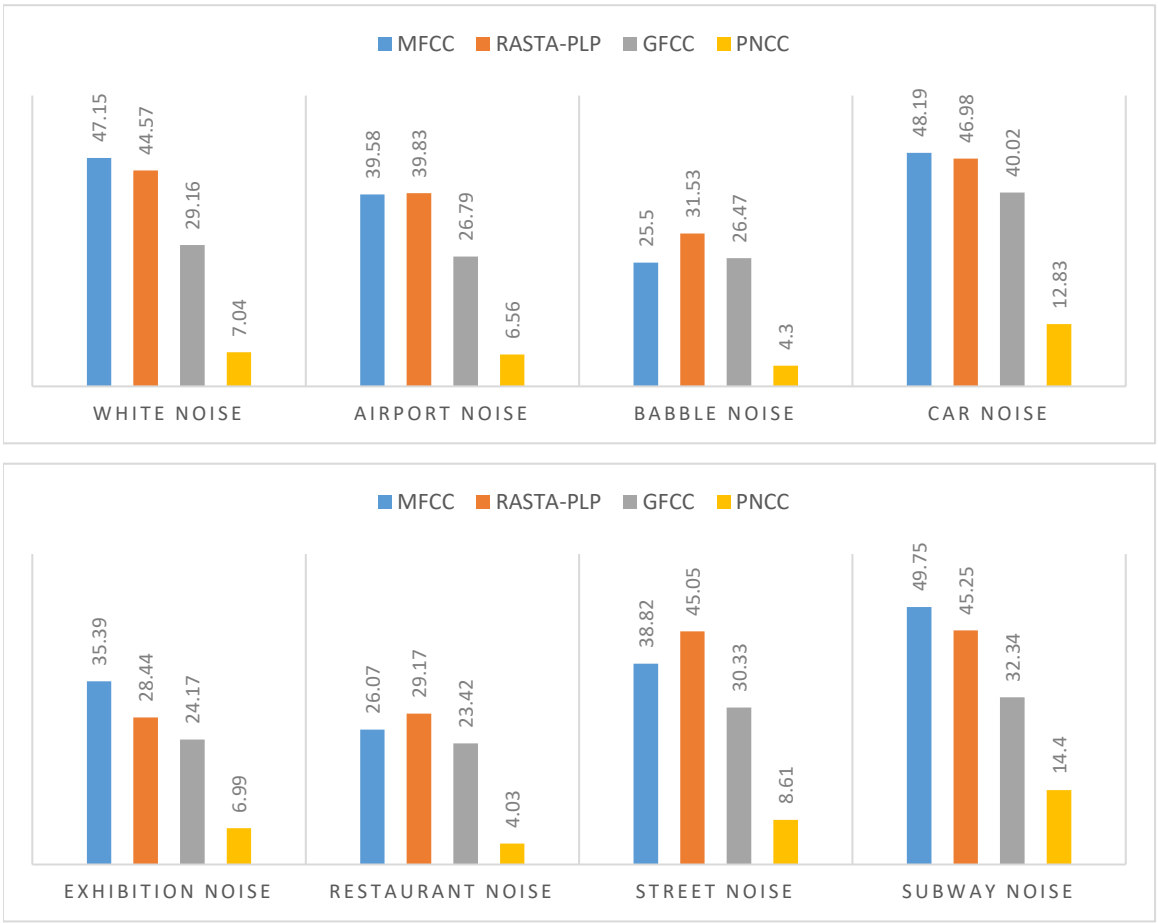


Figure 4-13: Percentage improvement rate for all types of noise at SNR 0dB

The bar charts in Figure (4-14) demonstrate the percentage improvement rate of the proposed method compared to the state-of-art methods at 5 dB. As shown in the figure, the proposed method still outcomes the other methods in the presence of white and environmental noise. The overall improvement in percentage recognition rate for all the types of noise is obtained in the case of MFCC and RASTA-PLP methods, alternately. Followed by GFCC method and lastly PNCC method. The highest recognition rate compared to MFCC, RASTA-PLP and PNCC methods is determined in the Subway noise condition. The proposed method is better than MFCC method by 55.72% while it is better than RASTA-PLP method by 47.87% while it is better than PNCC method by 8.16%. Differently, the highest recognition rate compared to GFCC method is determined in the case of Exhibition noise. The proposed method is improved by 24.85% in the case of GFCC method.

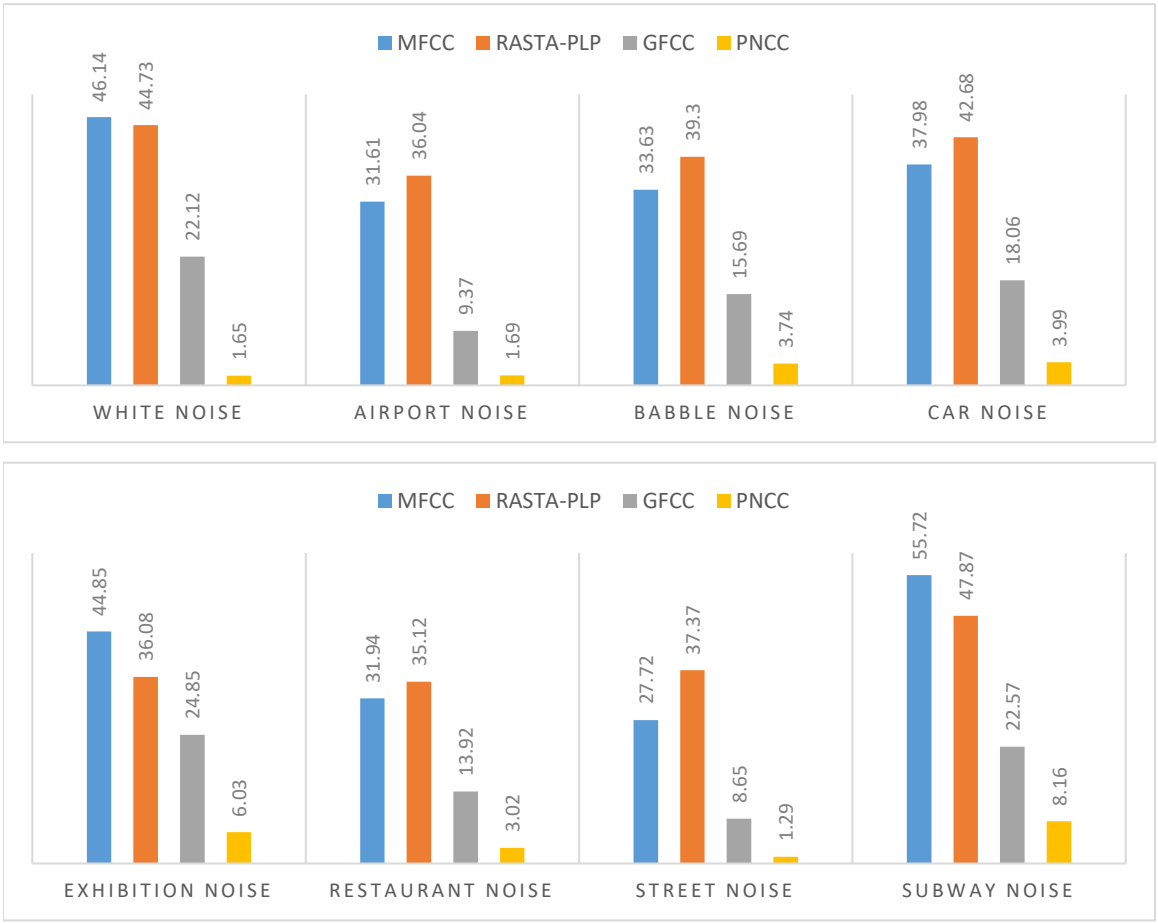


Figure 4-14: Percentage improvement rate for all types of noise at SNR 5dB

CHAPTER 5

CONCLUSION AND FUTURE WORK

5 Chapter Five: Conclusion and Future Work

5.1 Conclusion

This thesis seeks to improve the recognition accuracy in the noisy environment by improving the Automatic Speech Recognition systems robustness based on using techniques motivated by auditory processing. Two proposed methods have been developed in this thesis and they are classified as a feature-based approach.

The Hidden Markov Model was used as a machine learning tool which is based on the probabilistic or Bayesian model to calculate the probability for a sequence of events. The speech recognition systems are divided into two stages training stage and testing stage. In the training stage, the features were extracted from each word in the training datasets data set and by using a lexicon dictionary a set left-to-right Markov chains model are created for each word. Then using a multivariate Gaussian modeling technique and Baum-welch training algorithm, the Hidden Markov Model were created for each word and saved. Whereas in the testing stage, the features were extracted from each word in the testing datasets and the Viterbi decoder is used to match each word with the saved model to recognize it. Next, the performance of the system is evaluated in the terms of recognition accuracy.

In this thesis, the performance of the proposed methods are compared to the state-of-the-art techniques such as MFCC, RASTA-PLP, GFCC, and PNCC. The MFCC method, the speech waveform in analyzed in short spectral windows. Each window is filtered by a set of warped triangular filter banks to Mel-scale that models human auditory perception system. Then, the cepstral MFCC features are obtained by computing DCT of a log power spectrum. The RASTA-PLP system consists of a RASTA filtering technique combined with the PLP system. The PLP is another approach or method that models the human auditory system. It is based on three techniques that are derived from the physiology of hearing. The first technique is the critical-band spectral resolution, the second technique is the equal-loudness curve, and the last technique is the intensity-loudness power law. RASTA process is a noise filtering technique that used in suppressing is steady background noise. The GFCC this system is similar to MFCC system but it uses a Gammatone filtering technique instead of triangular Mel-filtering to emulate the human auditory perception system. The last method is PNCC. In this method, several techniques were used to suppress the noise. Such as, Asymmetric Nonlinear filtering to estimate the level of the acoustical background noise for each time frame and frequency bin. Mean Power Normalization that works as an automatic gain control to decreases the impact of amplitude variation of the

incoming acoustic wave. The Power Function nonlinearity curve which fit the relation between perceived signal amplitude in a given frequency channel and the related response of the auditory processing model.

In addition to feature extraction methods, a Feature Moment Normalization technique was added by applying Cepstral Mean Normalization to move all the extracted features to have a zero mean and additional speech features such as energy feature of each frame and dynamic features which represent the velocity and acceleration variation between frames.

The first proposed method is based on standard MFCC system and it is designed to improve the speech recognition system in the presence of white Gaussian noise. In this system, the adaptive time-frequency mapping technique is used to estimate which part of the uttered word is highly affected by noise and weighting it. The effect of time-frequency adaptive noise rejection has dramatically improved the recognition at low SNRs. The performance of the proposed system was examined by using TIDIGITS database. The experimental results show that the proposed method outperforms other methods in terms of recognition accuracy. The highest enhancement in recognition accuracy is obtained at SNR 0dB. It improved by 34.45% in comparison to of standard MFCC system and by 20.37% in the case of RASTA-PLP system.

The second proposed method is based on modifying the PNCC system to improve the performance in the presence of white and background environmental noise. The Medium Time average filtering is applied by running wide average filter along the Gammatone channels and then applying Channel Bias minimizing to decrease the bias along each channel which is produced from smoothing the noise. The effect of these techniques enormously improved the recognition for white noise and seven different types of environmental noise at low SNRs. The performance of the proposed system was also examined by using TIDIGITS database. The experimental results show that the proposed method outperforms other methods in terms of recognition accuracy. The highest improvement in recognition rate in comparison to MFCC and RASTA-PLP methods are obtained in Subway noise condition at SNR 5dB. The recognition rate improved by 55.72% more than MFCC method and 47.87% more than RASTA-PLP method. However, the highest improvement in recognition rate in comparison to GFCC and PNCC methods are obtained in the case of Car noise at SNR 0dB and -5dB, respectively. In this case, the recognition rate improved by 40.02% in comparison to GFCC method and 19.51% in comparison to PNCC method.

5.2 Future Work

Many different adaptations, tests, and experiments are suggested in the future work. They concern the deeper analysis of specific mechanisms, try different methods to enhance the proposed systems performance or to evaluate the proposed systems behavior at more different conditions such as:

- Modify the adaptive masking technique in the first proposed system to suppress more types of environmental noise.
- The performance of both proposed systems can be examined with larger vocabulary datasets and different languages datasets.
- The performance of the systems needs also to be evaluated with more noisy conditions. For instance, reverberation noise effect, colored noises, and mixtures of environmental noises.
- The systems need to be tested in real-life speech recognition applications such as mobile applications.

References

[1] F. Jelinek, "Speech recognition by statistical methods," *Proceedings of the IEEE,* vol. 64, pp. 532-556, 1976.

[2] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 651-654, 1988.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 22, pp. 745-777, 2014.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

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

[6] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language,* vol. 24, pp. 77-93, 2010.

[7] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on speech and audio processing,* vol. 7, pp. 55-69, 1999.

[8] A. M. A. Ali, J. Van der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Transactions on Speech and Audio Processing,* vol. 10, pp. 279-292, 2002.

[9] U. H. Yapanel and J. H. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication,* vol. 50, pp. 142-152, 2008.

[10] A. Fazel and S. Chakrabartty, "Sparse auditory reproducing kernel (SPARK) features for noise-robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, pp. 1362-1371, 2012.

[11] N. Moritz, M. R. Schädler, K. Adiloglu, B. T. Meyer, T. Jürgens, T. Gerkmann, *et al.*, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," *Proc. of CHiME,* pp. 1-6, 2013.

[12] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, pp. 1315-1329, 2016.

[13] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach* vol. 247: Springer Science & Business Media, 2012.

[14] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, pp. 1635-1638, 2000.

[15] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, pp. IV-757-IV-760, 2007.

[16] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[17] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 14-22, 2012.

[18] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, pp. 1304-1312, 1974.

[19] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, pp. 733-736, 1998.

[20] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, pp. I-656-I-659 vol. 1, 2003.

[21] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, pp. 113-120, 1979.

[22] J. H. Martin and D. Jurafsky, "Speech and language processing," *International Edition*, vol. 710, 2000.

[23] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, p. 126, 1998.

[24] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The annals of mathematical statistics*, vol. 37, pp. 1554-1563, 1966.

[25] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc*, vol. 73, pp. 360-363, 1967.

[26] L. E. Baum, "An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.

[27] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, pp. 268-278, 1973.

[28] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Electrical and Computer Engineering, 1995. Canadian Conference on*, 1995, pp. 1062-1065.

[29] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in *ISMIR*, 2000.

[30] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory physiology and perception*, vol. 83, pp. 429-446, 1992.

[31] B. C. Moore, *An introduction to the psychology of hearing*: Brill, 2012.

[32] R. D. Patterson, "Auditory filters and excitation patterns as representations of frequency resolution," *Frequency Selectivity in Hear-ing*, pp. 123-177, 1986.

[33] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH*, 2009, pp. 28-31.

[34] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *ICASSP*, 2010, pp. 4574-4577.

[35] C. Lemyre, M. Jelinek, and R. Lefebvre, "New approach to voiced onset detection in speech signal and its application for frame error concealment," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4757-4760.

[36] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, pp. 556-565, 2009.

[37] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoustics Research Letters Online*, vol. 2, pp. 91-96, 2001.

[38] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *The Journal of the Acoustical Society of America*, vol. 109, pp. 648-670, 2001.

[39] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Johns Hopkins University, 2010.

[40] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, 2005.

[41] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, 2001, pp. 103-106.

[42] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 188-193.

[43] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 289-292.

[44] M. Athineos, H. Hermansky, and D. P. Ellis, "LP-TRAP: Linear predictive temporal patterns," IDIAP2004.

[45] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681-684, 2008.

[46] R. Leonard, "A database for speaker-independent digit recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, 1984, pp. 328-331.

[47] D. Ellis. (2016, 17 Sept). *PLP and RASTA (and MFCC, and inversion) in MATLAB Using melfcc.m and invmelfcc.m*. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/

[48] C. M. University. (2016, 17 Sept). *CMU dictionary*. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict