Menoufia University
Faculty of Electronic Engineering
Dept. of Electronics and Electrical
Communication Engineering

Utilization of Deep Learning Techniques for Speech Signal Analysis

A Thesis

Submitted for the Degree of Ph. D. in Engineering Science,
Electronics and Electrical Communication Engineering
Signal Processing,
Department of Electronics and Electrical Communication Engineering

By

Eng. Samia Abd EL-Moneim Kabel

B. Sc. In Electronic Engineering, Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

M. Sc. In Electronic Engineering, Department of Electronics and Electrical Communication Engineering,
Faculty of Electronic Engineering, Menoufia University

Supervisors

Prof. Mohamed Mohamed Abd El-Salam Nassar
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Moawd Ibrahim Dessouky
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Nabil Abd El-Wahed Ismail
Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University

Assoc. Prof. Adel Shaker El-Fishawy
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

2020

Menoufia University
Faculty of Electronic Engineering
Dept. of Electronics and Electrical
Communication Engineering

Utilization of Deep Learning Techniques for Speech Signal Analysis

A Thesis

Submitted for the Degree of Ph. D. in Engineering Science,
Electronics and Electrical Communication Engineering
Signal Processing,
Department of Electronics and Electrical Communication Engineering

By

Eng. Samia Abd EL-Moneim Kabel

B. Sc. In Electronic Engineering, Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

M. Sc. In Electronic Engineering, Department of Electronics and Electrical Communication Engineering,
Faculty of Electronic Engineering, Menoufia University

Supervisors

Prof. Mohamed Mohamed Nassar ( )
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Moawd Ibrahim Dessouky ( )
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Nabil Abd El-Wahed Ismail ( )
Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University

Assoc. Prof. Adel Shaker El-Fishawy ( )
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

2020

Menoufia University
Faculty of Electronic Engineering
Dept. of Electronics and Electrical
Communication Engineering

Utilization of Deep Learning Techniques for Speech Signal Analysis

A Thesis

Submitted for the Degree of Ph. D. in Engineering Science,
Electronics and Electrical Communication Engineering
Signal Processing,
Department of Electronics and Electrical Communication Engineering

By

Eng. Samia Abd EL-Moneim Kabel

B.SC. in Electronics and Electrical Communication Engineering,
Faculty of Electronic Engineering, Menoufia University
M. Sc. in Electronics and Electrical Communication Engineering,
Faculty of Electronic Engineering, Menoufia University

Approved by

Prof. Ashraf Abd El-Moneim Khalaf (   )
Department of Electrical Engineering, Faculty of Engineering,
Minia University

Prof. Moawd Ibrahim Dessouky (   )
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Nabil Abd El-Wahed Ismail (   )
Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University

Prof. Osama Fawzy Zahran (   )
Department of Electronics and Electrical Communication Engineering, Faculty of Electronic Engineering, Menoufia University

2020

بسم الله الرحمن الرحيم

ACKNOWLEDGEMENTS

Thanks to ALLAH who give me this opportunity to do this kind of scientific work, and I hope that ALLAH accepts this work as a good deed.

In the development, presentation and compilation, of our work, I would like to give our sincerest gratitude to all those people who provided their guidance and support throughout. Without their support this dissertation would not have been possible.

I would like to thank all those who helped me with their knowledge and experience. I will always appreciate their efforts.

Thanks to my supervisors **Prof. Mohamed Abd El-Salam Nassar, Prof. Moawd Ibrahim Dessouky**, **Prof. Nabil Abd El-Wahed Ismail,** and **Prof. Adel Shaker El-Fishawy** for their supervision, valuable advice, fruitful scientific discussions, support, and the time and effort they gave me.

Also, I would like to give my sincere thanks to **Prof. Fathi Abd El-Samie**, who gave me the chance to work in the area of signal processing and deep learning, for his valuable proposals, respected discussions, support, time and effort he gave me to find my research path and the ideas that lead to the effective solutions. It was a great pleasure and honor to work with him.

At last but not least, my genuine appreciation and gratitude to my **father**, my **mother** (ALLAH forgive her), my **mother in law** (ALLAH forgive her), my **husband**, my sister (**Hanaa**), my son (**Anas)** and my daughter **(Maria)** for their help, and patience during the preparation of this work and unlimited support. Which is the source of motivation and gladness for me.

LIST OF PUBLICATIONS

[1] Samia. A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Text-independent Speaker Recognition Using LSTM-RNN and Speech Enhancement”, Multi Media Tools and Applications 79(33), pp. 24013–24028, June 2020.

[2] Samia. A. EL-Moneim, El-Sayed M. EL-Rabaie, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Speaker Recognition Based on Pre-processing Approaches”, International Journal of Speech Technology (IJST) 23, Springer, pp. 435–442, DOI.org/10.1007/s10772-019-09659-w, March 2020.

[3] Samia A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Quality Evaluation of Reverberant Speech Based on Deep Learning”, Menoufia Journal of Electronic Engineering Research (MJEER), Article 11, Vol. 29, Issue 2, pp. 126-132, Aug. 2019.

[4] Samia. A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Effect of Reverberation Phenomena on Text-independent Speaker Recognition Based Deep Learning”, International Conference on Electronic Engineering (ICEE), Vol. 1, pp. 19-23, Dec. 2019.

[5] Samia. A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Cancellable Template for Speaker Recognition Based on Spectrogram Patch Selection and Deep Convolutional Neural Network”, accepted for publication in International Journal of Speech Technology (IJST).

[6] Samia. A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Text-dependent and Text-independent Speaker Recognition of Reverberant Speech Based on CNN”, accepted for publication in International Journal of Speech Technology (IJST).

SUBMITTED PAPERS

[1] Samia Abd El-Moneim, Hossam Hammam, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Waleed El-Shafai, Atef Abu El-Azm, Mohammed El-Halwany, Fathi E. Abd El-Samie, “Effect of Interference on Text-independent Speaker Recognition Based on Deep Learning”, Multi Media Tools and Applications, Major revision.

[2] Samia. A. EL-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Fathi E. Abd El-Samie, “Performance Enhancement of Text-independent Speaker Recognition in Noisy and Reverberation Condition Using Radon Transform with Deep Learning”, International Journal of Speech Technology (IJST), Major revision.

Abstract

This thesis is mainly concerned with text-independent Speaker Recognition (SR). Generally, the Automatic Speaker Recognition (ASR) system can be classified into two main categories: text-dependent SR and text-independent SR. In text-dependent SR, all speakers are committed to use the same sentence in both training and testing phases. On the other hand, in text-independent SR, speakers are free to use any sentences in the training and testing phases. The SR process in general depends on the extraction of features from the speech signals. The text-independent SR task is harder to implement than the text-dependent SR task. Two proposed approach are introduced in this thesis for text-independent SR.

The first proposal depends on extracting features and utilization of Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to identify the speakers. The utilized features are Mel Frequency Cepstral Coefficients (MFCCs), spectrum magnitude bins, and log spectrum magnitude bins. The second proposal depends on the generation of spectrogram images from the speech signal patches. These spectrogram images are utilized in the classification process with a Convolutional Neural Network (CNN).

The reverberation is a severe effect that exists in closed rooms. A proposed speech classification system is introduced to classify the speech signals into reverberant or not using the LSTM-RNN and the CNN. The effects of noise, reverberation, and interference are considered in this study. Moreover, speech enhancement techniques such as spectral subtraction and wavelet denoising are considered in this thesis to enhance the performance of the SR process. These enhancement methods are used as a pre-processing steps prior to the ASR system. In addition, Radon Transform (RT) is used for better representation of speech signals in the presence of noise as it is robust to the noise effect. The Radon projection of the spectrogram of speech signals is obtained at different orientation or angles, A DCT is then taken after applying Radon projection. The performance of the ASR system with Radon features is compared to that with MFCCs and spectrum. Also, the effect of interference on the ASR system is studied. The interference effect is cancelled with a signal separation algorithm that is used as a pre-processing step prior to the ASR system to boost its performance. For pattern security of the SR system, cancellable SR is presented in this thesis with an approach that depends on spectrogram patch selection based on a user-specific key. The Cancellable pattern is used to protect the user privacy and increase the its security.

Simulation results prove the high efficiency of the proposed approaches for text-independent SR with the enhancement methods, Radon based features and blind signal separation. Also, the results reveal that, the suggested cancellable approach is practical, and satisfies the desired criteria of renewability, security [which means that the template can be changed if it is compromised], and high performance [which is near to the performance of the system with the original template].

Table of Contents

CHAPTER 1 Introduction

CHAPTER 2 Literature Review

CHAPTER 3 Mathematical Models

CHAPTER 4 Proposed Speaker Recognition System Based on Deep Learning

List of Abbreviations

ASR	Automatic Speaker Recognition
ATM	Automated Teller Machine
AWGN	Additive White Gaussian Noise
CNN	Convolutional Neural Network
CNV	Convolutional Layer
CL	Classification Layer
DL	Deep Learning
DCT	Discrete Cosine Transform
DST	Discrete Sine Transform
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
FC	Fully Connected
FT	Fourier Transform
GMM	Gaussian Mixer Model
HMM	Hidden Markov Model
IDFT	Inverse Discrete Fourier Transform
LSTM	Long Short-Term Memory
MFCCs	Mel Frequency Cepstrum Coeficients
PIN	Personal Identification Number
RIR	Reverberant Impulse Response
PLP	Perceptual Linear Prediction
RNN	Recurrent Neural Network

RT	Radon Transform
ReLU	Rectified Linear Unit
SR	Speaker Recognition
SS	Spectral Subtraction
STFT	Short-Time Fourier Transform
T	Threshold
TFR	Time Frequency Representation
VADCS	Voice Activated Device Controls
VQ	Vector Quantization
Wden	Wavelet Denoising

List of Symbols

b	Bias
C_g	MFCCs
c	Cell
c_t	Current cell memory
c_{t-1}	Previous cell memory
$e(n)$	Excitation source
f	Forget gate
$f(x,y)$	2-D signal
$f_{hard}(x)$	Hard thresholding
$f_{soft}(x)$	Soft thresholding
h_t	Current cell output
h_{t-1}	Previous cell output
$h(t)$	Room impulse response
$\eta(x, y$	White noise
i	Input gate
L	Window length
n	Time index
N	Number of speaker
o	Output gate
r	distance
R	Recurrent weight
$z(t)$	Reverberant signal

$R(r,\Theta)$	Radon transform
$s(n)$	Speech signal
$S(\tau,\omega)$	Speech spectrum
$v(n)$	Vocal tract impulse response
ω	Frequency index
$\mathbf{w}_c$	Cell weight
$\mathbf{w}_f$	Forget weight
$\mathbf{w}_i$	Input weight
$\mathbf{w}_o$	Output weight
$W(n-\tau)$	Window function
$x(n)$	Discrete signal
x_t	Network input
$y(n)$	Noisy speech
Θ	angle
$\Phi(x)$	ReLU activation function
$*$	Convolution
Θ	Multiplication of vector
$\Re$	Radon transform
$\delta(.)$	Dirac delta function
σ	Sigmoid activation function

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Overview

Biometrics is the most relevant means of authenticating and recognizing persons in a consistent and fast way over the use of distinctive anatomy. Biometrics allows an individual to be recognized and authenticated based on a set of recognizable and verifiable data, which are unique and specific to them. Biometrics is based on physiological or behavioral features to identify persons. There are several biometrics recognition modalities: face recognition, fingerprint recognition, iris recognition, voice, and signature recognition. In contrary to Personal Identification Numbers (PINs) and passwords, biometrics techniques are most appropriate relative to the user and they prevent unauthorized access or fake use of Automated teller machines (ATMs), time and attendance systems, cellular phones, smart cards, and desktop PCs. The identification based on biometric techniques avoids the necessity to remember a password. Automatic Speaker recognition (ASR) has been realized as a security scheme to control admission to the buildings or information. In addition to adding security, ASR also offers a mechanism to limit the remote access of a personal workstation to its owner or a set of registered users [1].

The human auditory system is able to identify speakers, normally or over the telephone by listening to their speech. Accomplishing this inherent skill is the main task of the ASR system. Similar to human listeners, ASR uses the speech signals to determine the speaker's personality. The ASR system performs the task of authenticating the users' identities using features extracted from their voices.. Voice is one of the biometric keys that will be used in this thesis. The ASR can be defined as the tool that uses the human voice to identify or verify the speaker's identity [2].

The speech articulation pronounced by a person such as "Good morning, how're you?" carries a large amount of information or message that can be extracted. Also, this message contains valuable information about the speaker's characteristics as identity, age, dialect, health, and mood state. An ASR system tries to imitate the way of human being recognition system. It is required for the ASR is to be familiar with the heard sound. This can be done by training, which means that the system has to know who is speaking. This is the registration stage, wherever a voice signature from every person is taken and stored in the database. After that, this voice is used in the training stage as a model of each speaker is created. The model is stored and it is an indication to a certain person identity. Finally, in the test stage, any speaker can present his voice to the system, and in the same manner, as in training, a model is created. The system matches the unknown model with the models that have been stored and a decision is taken about the identity of the speaker [3].

Voice biometric is a highly speaker reliant aspect, which varies from person to another based on individual physical traits, these physical traits such as pitch, tone, and intensity vary from speaker to another. Using many existing algorithm, speech signal analysis and measure became easier than before. For example the neural network, support vector machine, Gaussian Mixer Model (GMM),… etc. are used for pattern matching, also MFCCs, LPCs,… are used for features extraction. Since only a few speakers model parameters have to be stored, memory requirements are not high compared to 2D or 3D-based biometric keys as in iris or face recognition. All these aspects sort voice to be an influential biometric key to be used in many security applications [4, 5].

Although the voice is an actual suitable biometric, there are some demerits as mismatching between training and testing utterances. This mismatching can result due to the health conditions and microphone or channel used to carry voice. In this case, the ASR will not be competent to recognize the speaker. There are some algorithms that can diminish the influences of the microphone and the transmission channel over the speech signal. In this thesis, our study is carried out in an acoustic environment place.

In this introduction, some main concepts and a general idea of the ASR system are presented. The types of ASR systems are discussed. Also, degradations that affect the ASR are introduced. In addition, different applications of this technology will be discussed.

1.2 Main Concepts of Speaker Recognition System

Voice processing have several applications such as speech recognition, speaker recognition, speech coding, speech synthesis, and speech enhancement. The speech and speaker recognition are two distinctive applications, each has its own set of technologies and uses. The speech recognition is the process of extracting usable linguistic information from a speech signal in support of human-machine communication by voice. In other words, speech recognition is the process of recognizing individual words or phrases from human speech. On the other hand, speaker recognition is the process of recognizing the person identity from his spoken words. In summary, speech recognition is more related to recognizing words or phrases, while speaker recognition is related to detecting who is the speaker [6].

The ASR includes two main phases: the training and the testing phases. The training phase aims to get a voiceprint of a given speaker. This can be done by asking the registered speaker to say a specific phrase that can be captured by a microphone. The speech signal is pre-processed by an A/D and converted into frames. The next step is feature extraction, which is considered as a dimensionality reduction of the speech signal. It converts the speech into a low-dimensional vector that is suitable for the pattern matching process. Feature extraction is the main block in the ASR system, since it captures the speaker-dependent features. The last step is the pattern matching, which operates in two modes; training mode and testing mode as shown in Fig. 1.1. In the training mode, the feature vectors obtained from a known speaker are used to make a model for this speaker and store it in a database. In the testing phase, the model of the unknown speaker is compared to the already existing models in the database by a pattern matching algorithm [6,7].

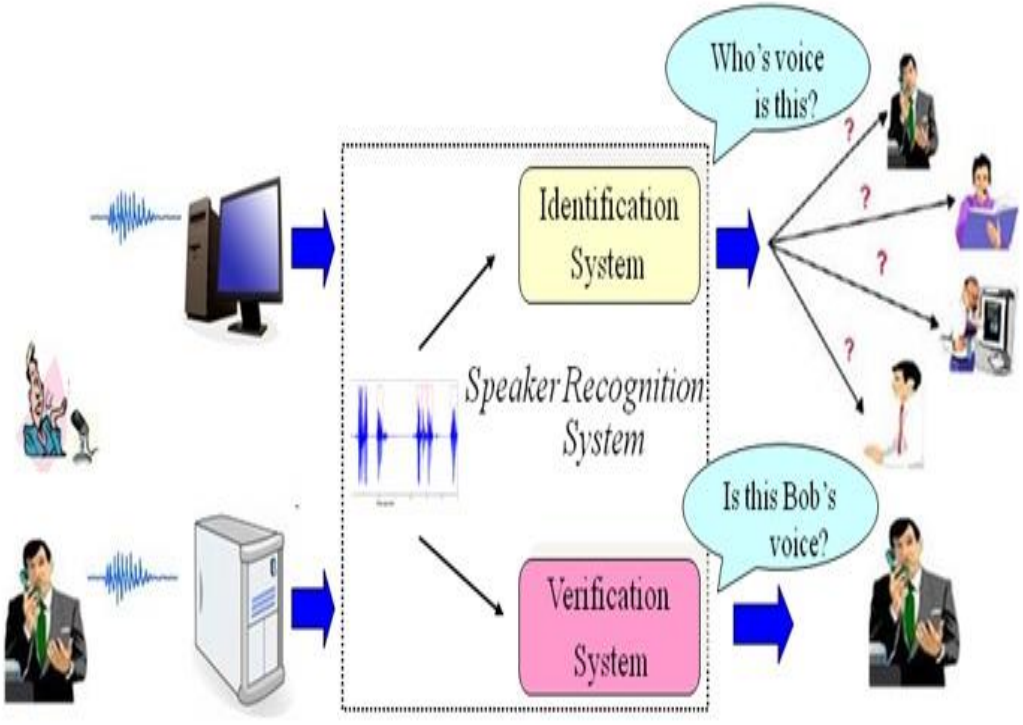


Fig.1.1 Speaker recognition system [6].

The testing phase is similar to the training phase, except that the speaker is unknown. Also, the pattern matching algorithm compares the unknown speaker's model with the models that are already obtained from the enrollment stage to get the best match. The matching results are obtained as a score, and according to this score, a decision is made about the identity of the speaker.

1.3 Speaker Recognition Sub-categories

The ASR system has two stages: registration and verification. Through registration, the speaker's voice is recorded and generally, several features are extracted to make a model for each speaker. In the verification phase, a speech sample or "utterance" is compared against a previously created models. ASR systems can be categorized into three different categories. Firstly, they can be divided into speaker recognition or speaker authentication. Secondly, they can be text-dependent or text-independent systems. In a text-dependent system, the text is the same during the enrollment and verification phases. In text-independent systems, the text during enrollment and test is different. Finally, it is possible to sort into open-set or closed-set systems. In the next sections, all these categories will be described [8].

1.3.1 Speaker Authentication and Speaker Recognition

The goal of authentication or verification is to verify the speaker's claimed identity based on his or her utterance. Only a yes/no decision is taken as the user provides his identity to the system. The speech features of the claimed speaker are compared to the model attached to the identity provided by the user. So, a single comparison is performed and this results in a single decision. The speaker recognition is different from authentication, as speaker recognition implies both recognition and authentication. The speaker utters a specific phrase and the system selects which speaker model matches the speech input. For N speakers, there will be N comparisons. For each comparison, a score is created, and so the system indicates the identity close to the most likely speaker model. It is obvious that speaker recognition is more complicated than speaker verification, and so the obtained results from recognition are more poorer than those of verification [8,9].

1.3.2 Text-dependent and Text-independent Speaker Recognition

The ASR system can be categorized based on the sentence or text used for recognition into text-dependent speaker recognition and text-independent speaker recognition. In text-dependent speaker recognition, the speaker should say the same sentence in both training and testing. This means that there is a particular word or phrase that the speaker should say to be recognized. On the other hand, the text-independent speaker recognition does not take the message contained in the speech input into consideration. So, this system can recognize the speaker from any spoken phrase and it is not limited to recognition of speakers based on the same sentences stored in the database. Training and testing are based on any utterance produced by the speaker. This type adds flexibility to the system, and enlarges its utility as it avoids the necessity to remember any specific password [10,11].

1.3.3 Closed-set and Open-set Speaker Recognition

Another categorization of the ASR system is to open-set and closed-set speaker recognition. In closed-set speaker recognition, the unknown speaker is one of the registered speakers. In other words, the speaker who is attempting to access the system belongs to the set of known speakers. On the other hand, in the open-set speaker recognition, the unknown speaker can be either registered or unregistered. This type is more complex and matches well the conditions in real life [11].

1.4 Degradations Effects on Speaker Recognition Systems

The ASR system can commonly behave well in perfect conditions without any source of distortion. Noise, reverberation, and interference are listed at the top of ASR challenges. These degradations of speech cause the ASR systems to perform poorly [12]. Since these degradations cause a severe mismatch between training and test utterances, they reduce the ASR system performance. The effects of these different impairments on the ASR system performance are considered in this dissertation. The noise sources may vary, and they are not limited to mismatches in the microphone or recording apparatus. Reverberation is another source of distortion that affects the speech in closed rooms.

Reverberation and echo are two distinct concepts. Echo is the reflected signal from the original speech signal, when the original speech travels a considerable distance, while reverberation is the combination of the original speech with multiple reflected versions. This may be displayed as additional sound sources added to the system. Reverberation can be defined as the persistence of sound in space after a sound source has been stopped. The distortion of the recorded signal is dependent upon the reverberation and absorption characteristics of the room, as well as the objects within the room. It can be modeled as an impulse response of a linear system. The impulse response provides a model of all possible paths that the sound sources take to arrive at the microphones. High reverberation makes the room sound loud and noisy, which reduces the speech quality and intelligibility, and hence reduces the ASR performance [13].

Interference is an important issue that must be taken into consideration. Interference is a mixing of speech signals recorded in a typical acoustic environment place. For the speech signals recorded in a typical room with an array of microphones, each microphone receives a direct copy of the sound source with some propagation delay based on the location of both the sources and the microphones [14]. In addition, several reflected and modified (attenuated and delayed) copies of the sound are received. Blind signal separation of mixed speech is a pre-processing method that can be used to separate the mixed signals to enhance the ASR performance.

1.5 Applications of Speaker Recognition Systems

ASR can be realized in many security areas, since it is a biometric security tool. Currently, banks, shops, or other kinds of business companies permit their clients to carry out different processes by telephone. This is very easy for the user, since nearly any kind of shopping or bank contracts can be performed from any place in the world. At this time, the problem of security raises. A non-authorized user can displace an authorized user’s identity very simply. The ASR system can diminish considerably the possibility that an impostor displaces an actual user identity over the telephone line. As the voice is enough to verify the user’s identity, there are different telephone services such as voice mail.

ASR can also be well-practiced in situations, where the speaker’s existence is needed. The Voice-Activated Device Control (VADC) is a related application in this area. The VADCs are devices such as doors, and locks are activated by voice without the need for a key or remote control. Computers, networks, or credit card transaction access can also be accomplished by voice instead of a consumer name or a PIN code. Law enforcement is another application, and it has been used in criminal and forensic inquiries to investigate the recorded speech offered as a proof in the trials [15]. In conclusion, it can be seen that ASR is realistic in many areas, and not all of them are concerned with security matters. The ASR is a cheap technology that can be applied in many cases.

1.6 Problem Definition

Most studies in ASR applications are considered as text-dependent applications with an ideal scenario for training and testing without noise, reverberation or interference. In real cases, some sources of degradation may exist. So, in this dissertation, we consider text-independent SR in the realistic case of degradation, Fig 1.2 shows the all proposed approach that introduced in this thesis.

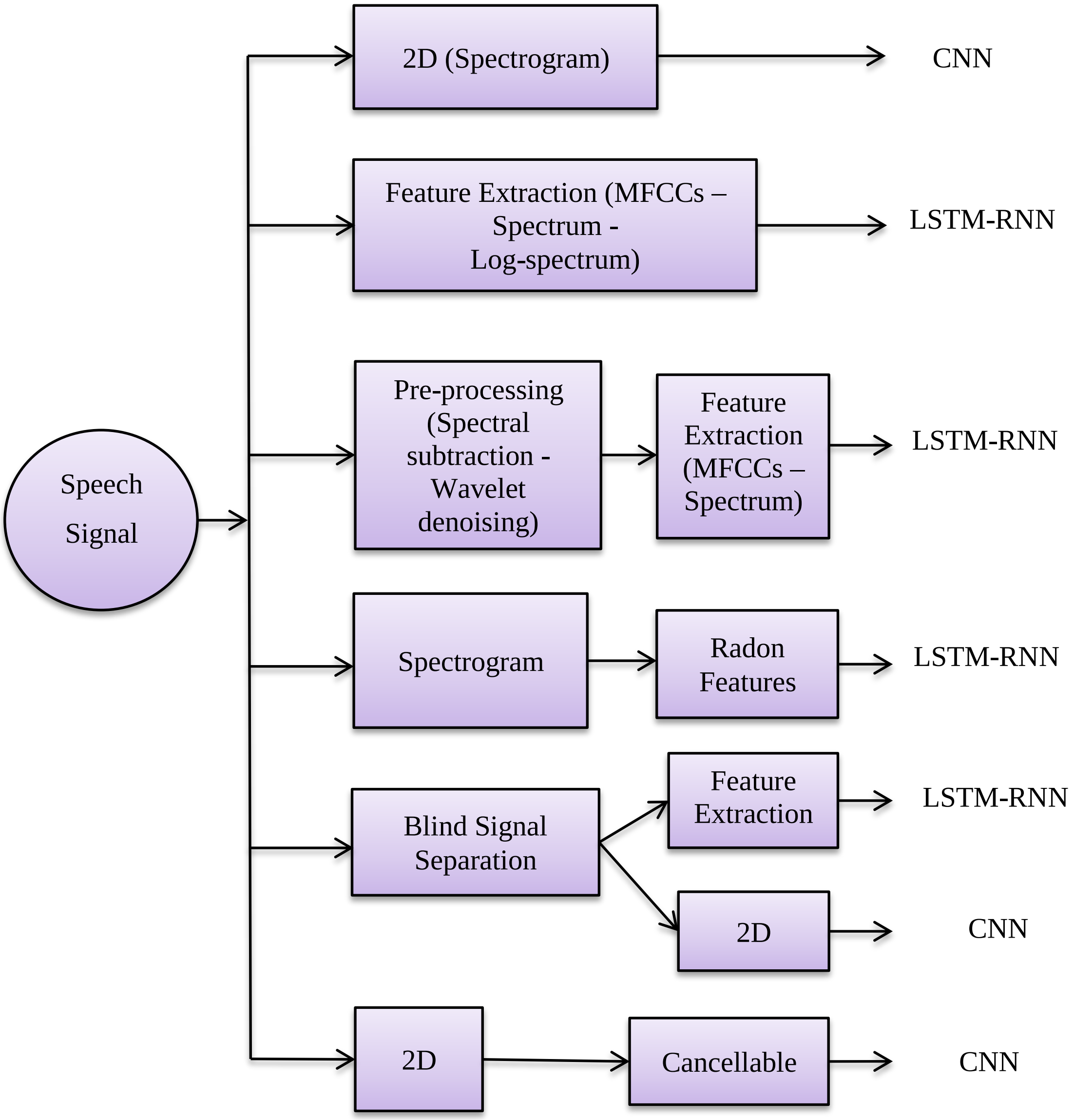


Fig 1.2 Flowchart of the proposed approaches.

The effects of noise, reverberation, and interference are considered in this thesis. Moreover, speech enhancement techniques such as spectral subtraction and wavelet denoising are considered to enhance the performance of the speaker recognition process. In addition, RT is used for better representation of speech signals in the presence of noise as it is robust to the noise effect. Also, the interference effect is cancelled with a signal separation algorithm. For pattern security of the ASR system, cancelable ASR is presented in this thesis with an approach that depends on spectrogram patch selection based on a user-specific key.

1.7 Thesis Objectives

The focus of this thesis is on text-independent speaker recognition system. Text-independent speaker recognition is highly flexible and can be applied in several applications. Text-independent speaker recognition is the only approach that can be done entirely unobtrusively, they need quite little cooperation from the individual .It is the most difficult type of ASR. While the text-dependent system is vulnerable to recordings, so it may need password-reset procedures if passwords are lost, stolen, and forgotten. An overview of the ASR systems has been shown in the previous section. Also, different types of ASR are presented. Degradations that affect ASR performance are also sorted such as noise, reverberation, and interference. All these degradations are explained and their influence on the ASR are shown in the following chapters. Some applications of ASR are also mentioned in the previous section.

The main objectives of the thesis are summarized in the following points:

- ❖ Presenting a review study of speech signals and the speech production model.

- ❖ Studying the fundamentals of the text-independent speaker recognition systems.
- ❖ Studying the traditional techniques of feature extraction.
- ❖ Investigation of different degradation effects on the ASR and presenting some enhancement methods to reduce these degradations.
- ❖ Utilization of DL models with ASR.
- ❖ Utilization of pre-processing techniques before ASR with DL.
- ❖ Utilization of RT for feature extraction to obtain robust features from speech signals in noisy environments.
- ❖ Studying the effect of interference on the ASR systems and utilization of signal separation techniques to reduce this effect.
- ❖ Implementation of cancellable ASR systems.

1.8 Thesis Organization

This thesis, describes the principles of speech production and perception. Furthermore, text-independent speaker recognition systems are proposed with Deep Learning (DL) technology, described and tested. The results of these tests are presented in the fourth chapter of the thesis. The effects of noise, reverberation, and interference are considered in this study. Some enhancement techniques are used to enhance the performance of the ASR. In addition, Radon Transform (RT) is used for better representation of speech signals in the presence of noise as it is robust to the noise effect. The interference effect is cancelled with a signal separation algorithm. A new conception is addressed to increase the security of the ASR systems based on a cancellable biometric technology. The Thesis is organized as follows:

Chapter 2 gives a literature review of the speech production model and ASR system theory, stages, and methods are described. DL technology and its different models are also presented. In **Chapter 3**, all mathematical models of the used methods are presented. The proposed ASR systems based on DL models are presented and the effect of different degradations on their performance are given in **Chapter 4**. A new concept is addressed in **Chapter 5** to increase the ASR security based on cancelable templates. The performance of the ASR with cancelable templates is studied, and the results reveal that, the system satisfies the required criteria as renewability, security, and high recognition performance. The conclusion is presented in **Chapter 6**.

CHAPTER 2

LITERATURE REVIEW

2.1 Synthesis and Characteristics of Speech Signals

The speech signal is an acoustic wave. It is the ancient way of communication between people. The birth of speech begins from the brain as a thought process, and then this thought is transformed into linguistic form. Speech is the outcome of a composite process carried out in the speaker's respiratory system. The main components included in this process are lungs, larynx and vocal tract. Every element presents some speaker-dependent information in the speech signal, and it also contributes somehow to the final speech signal. The speech is produced as a result of the excitation of the vocal tract.

Lungs are the power source that provides energy to the rest of the elements in the speech production model. Since lungs are responsible for airflow to larynxes, the larynxes modulate the airflow generating a periodic or a noisy airflow source and withdraw either of them into the vocal tract. The vocal tract forms the spectrum of the excitation source. The vocal tract entails nasal, oral, and pharynx cavities. The sound wave travels due to the change in air pressure at the lips to the listener in the form of a pressure wave. So, the speech production system is a high-level system [12].

In few words, there are three physiological apparatuses for speech generation: a sub-glottal constituent that comprises lungs and related respiratory muscles, larynges that comprise vocal folds, and vocal tract that embraces oral cavity, nasal cavity, epiglottis, tongue,.. etc. Figure 2.1 shows the basic speech production system. At the essence, the sound is shaped due to the vibration of the vocal tract. Consequently, the frequency at which the vocal folds vibrate is called the pitch frequency, which varies between persons due to the change in the size and the form of the vocal folds and glottis. Females have higher pitch frequencies than males because they have smaller larynx than males.

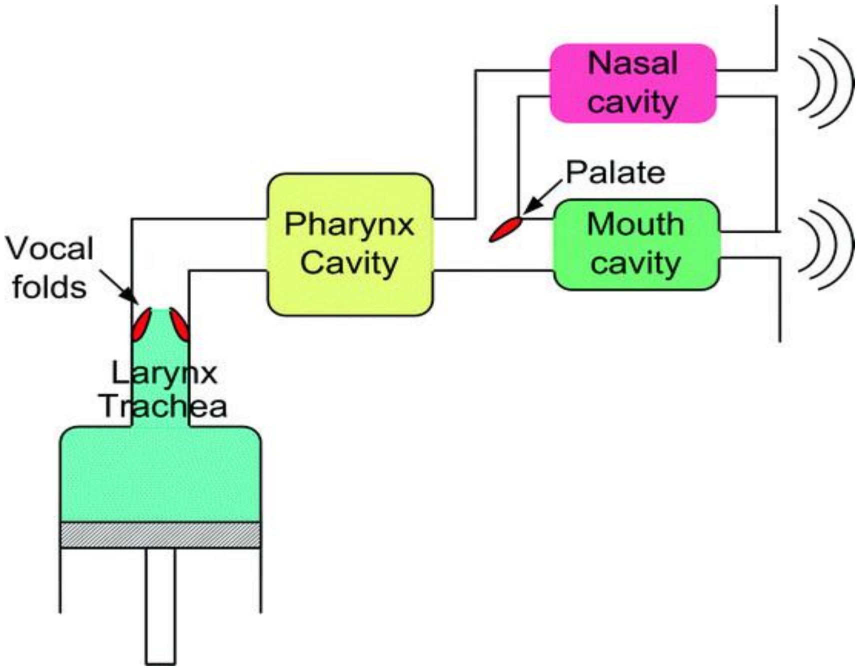


Fig. 2.1 Basic speech production system [16].

The pitch is modulated by the vocal tract. When the air flows through the vocal tract, it triggers it to vibrate at a certain frequency contingent on the length and diameter of the cavities in the vocal tract. This frequency is called the formant frequency. Each person's vocal tract utters each vowel with a dissimilar formant. So, the vocal tract can be referred to a filter, whose input is the pitch and its harmonics originating from the vocal folds. Concisely, every human has his specific filter (vocal tract) that differs from one to another [17].

2.2 Modeling of Speech Production System

As revealed above, the vocal tract can be represented by a filter, whose input is the excitation from vocal folds and output is the sound formed by the vocal tract. So, the conception speech production can be simplified by a source and a filter. The source behaves as the excitation source or the airflow from the larynx and the filter behaves as the vocal tract. Since the speech can be voiced or unvoiced, the pulse train (vibration of the vocal cord with a particular frequency called pitch) is the source for the voiced part and the random noise (air from the lungs) is the source of unvoiced speech as shown in Fig. 2.2. Tersely, the sound is demonstrated as a convolution of the excitation source and the vocal tract response shown in Eq. 2.1 [17].

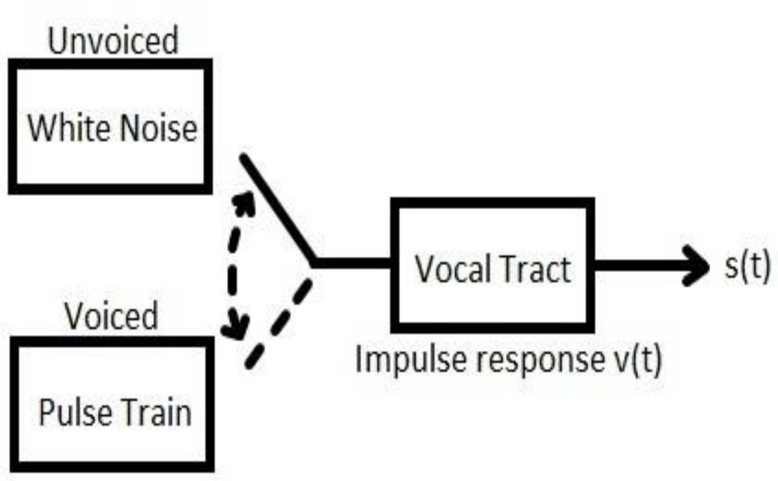


Fig. 2.2 Speech production model [16].

$$s(n) = e(n) * v(n) \tag{2.1}$$

where $e(n)$ is the excitation source and $v(n)$ is the vocal tract impulse response, and $*$ is the convolution sign. In the frequency domain, Eq. 2.1 becomes

$$S(K) = E(K)\,V(K) \tag{2.2}$$

where $V(K)$ is the Fourier Transform (FT) of $v(n)$ and $E(K)$ is the FT of $e(n)$ [12,16].

Because the speech signal has quasi-stationary nature, it is analyzed over a proper short period of time (20-30 milliseconds) in order to have relatively constant acoustic characteristics. The speech signal is divided into short segments called frames. A preset window length (usually 20-30 milliseconds) is moved along the signal with an overlapping (usually 30-50% of the window length) between the nearby frames [18, 19]. Overlapping is a pre-requisite to prevent loss or variations of data at the frame end. The process of disintegration of the signal into short segments is termed as framing and such segments are called windowed frames (or sometimes just frames) [18, 19].

2.3 Text-independent Speaker Recognition System

The ASR system is foremost in many applications. As mentioned in the previous chapter, the objective of this system is recognizing the speaker's identity and determining who he/she is. The ASR system imitates the human auditory system that hears the voice of the unknown speaker and makes a decision as to whom this voice is associated. The system labels each person to a certain voice signature. The ASR system can be classified according to the conveyed message into text-dependent speaker recognition system and text-independent speaker recognition system. In many ASR applications, it is possible to reduce the intra-speaker variability by necessitating the user to utter the test sentence that encompasses the same text or vocabulary as the training sentences. This is the situation of text-dependent speaker recognition systems. In text-dependent speaker recognition systems, the speaker is asked to utter a specified word or sentence in both training and testing. In this scenario, the system takes knowledge of the phrase the speaker would say. So, the system has a precognition of the spoken phrase, which leads to high recognition probability. On the other hand, in text-independent speaker recognition systems, there are no definite words the speaker is asked to say. The system does not take cognizance of the words said by the speaker, and the system has to determine the speaker from any expression. So, this system is considered more convenient because the user can speak freely to the system [18]**.** The ASR system comprises two essential phases; feature extraction and pattern matching as shown in Fig. 2.3.

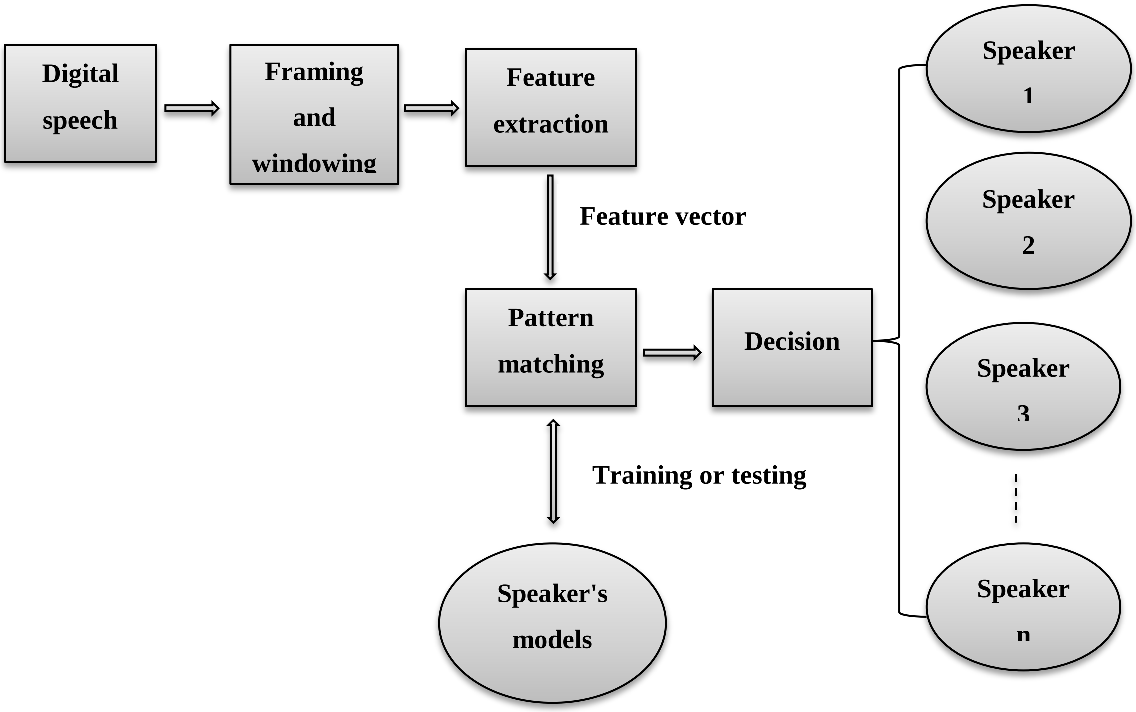


Fig. 2.3 Speaker recognition system [12].

In the feature extraction phase; speaker-reliant information is obtained from the speech signal. Most features are extracted from the physical traits in the vocal tract of the speaker. Pattern matching encompasses two important functionalities: training phase and testing phase. In the feature extraction process, a chain of feature vectors is obtained. This sequence is associated with a given set of speaker models using a pattern matching algorithm. This algorithm produces a likelihood score for each frame and speaker. With this information, a decision-making algorithm gives a decision about the unknown speaker identity [19].

2.3.1 Feature Extraction

The feature extraction process is the backbone of the ASR system. These features can be derived from physical traits or spectral information of speech signal called low-level features. In the ASR, there is a distinction between low-level and high-level information. The high-level information indicates the information such as dialect, pronunciation, speaking style, and subject style of the context. These features are only recognized and analyzed by individuals. The low-level features refer to information such as pitch, rhythm, tone, spectral magnitude, frequencies, and bandwidths of a person's voice. The ASR systems depend on the low-level features [12, 16].

The dissimilarity of features produced by different speakers is called inter-speaker variation. The inter-speaker variation is produced by different vocal characteristics of individuals and it offers valuable information for differentiating different speakers. The other type of variation is intra-speaker variation that arises when a speaker utters the same sentence, but cannot exactly reveal the utterance in the same manner from train to test. The intra-speaker variation comprises different speaking rates, emotional states of speakers, and speaking environments. The intra-speaker variation is the foremost issue that causes degradation in the ASR system performance. Consequently, it is necessary to choose the parameters that display lower intra-speaker variability, but higher inter-speaker variability. Since the features are the front end of the ASR, it is essential to comprehend right how the voice is produced and perceived by a human being [12].

There is a redundant information in the speech signal that is not appealing to the ASR system task. Feature extraction can eliminate most of this information, while emphasizing that is highly speaker-dependent. Also, reducing the speech signal dimension leads to a reduction in the quantity of information to be handled by the pattern matching algorithm, and hence a reduction of time consumption. There are different feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP),… etc. In this thesis, the features extraction methods used here are MFCCs, spectrum, and log-spectrum.

2.3.1.1 Mel Frequency Cepstral Coefficients (MFCCs)

The MFCCs are spectral features generally used in the ASR system. They are based on the speech processing similar to that takes place in the human ear. A few amount of computations is necessary with parameters such as linear predictive coding (LPC) coefficients. The steps of extracting the MFCCs are shown in Fig. 2.4.

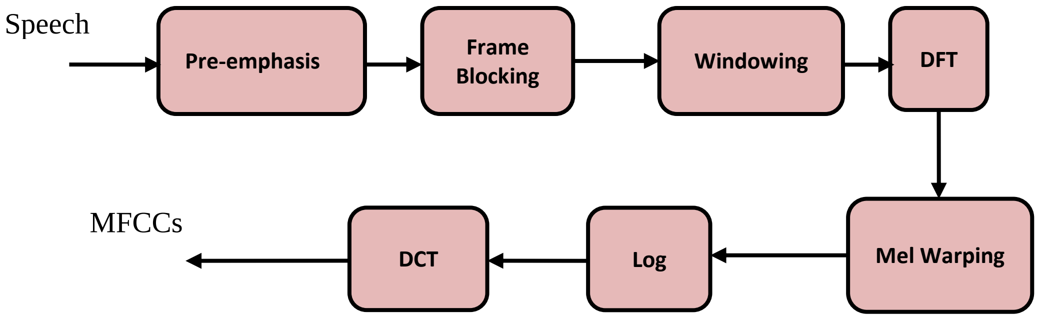


Fig. 2.4 MFCCs extraction [12]

Steps of estimating the MFCCs from a speech signal are:

- The speech signal is divided into short segments.
- The power spectrum of each segment is estimated.
- The Mel bank is applied on the power spectra, and the energy for every filter is summed [19].

$$Mel(f) = 2595 log_{10}(1 + \frac{f_{linear}}{700}) \tag{2.3}$$

- The logarithm of the filter bank energies is then taken.
- The Discrete Cosine Transform (DCT) of the logarithms of the filter bank energies is calculated.
- Coefficients from 1 to 13 are taken for every segment and others are discarded [19].

The MFCCs can be calculated as in Eq. 2.4 [19, 20].

$$C_g = \sqrt{\frac{2}{N_f}}\sum_{m=1}^{N_f} \log(\acute{S}\,(m))\cos(\frac{g\pi}{N_f}(m-0.5)) \qquad (2.4)$$

where C_g refers to the MFCCs, $\grave{S}(m)$ is the m_{th} Mel filter output, $j = 0,1,2,\dots J-1,$ g is the index of the MFCC, N_f is the number of Mel filters [19, 20].

2.3.1.2 Speech Signal Spectrum and Spectrogram

The spectrogram is a time-frequency representation of the signal, which includes the complete information of the signal in both spectral and time domains. Spectrogram is computed by applying Short Time Fourier Transform (STFT) on the signal through the segmentation of the signal into segments of fixed length, and then the application of a window with some overlap. The spectrogram is the squared magnitude of the STFT [21].

$$X(\tau,\omega) = STFT\{x(n)\} = \sum_{n=0}^{N-1} x(n)w(n-\tau)e^{-jn\omega} \qquad (2.5)$$

$$S(\tau,\omega) = |X(\tau,\omega)|^2 \qquad (2.6)$$

where $X(\tau,\omega)$ is the STFT of the speech signal $x(n)$, $w(n)$ is the window and $S(\tau,\omega)$ is the spectrogram. The spectrum can be extracted as a slice of the spectrogram. It can be seen that the STFT is a function of two variables, namely a time index τ and a frequency index ω, and hence a time-frequency representation is obtained. In understanding of speech signals, it is beneficial to see how the power spectrum changes over time and this makes the spectrogram a powerful tool for this task. The spectrogram of a speech signal is shown in Fig. 2.5. It is seen that, this is a function of both time and frequency. It should be noted that, the window length affects the appearance of the spectrogram. A long-duration window gives a better frequency resolution, but a poor time resolution [22, 23].

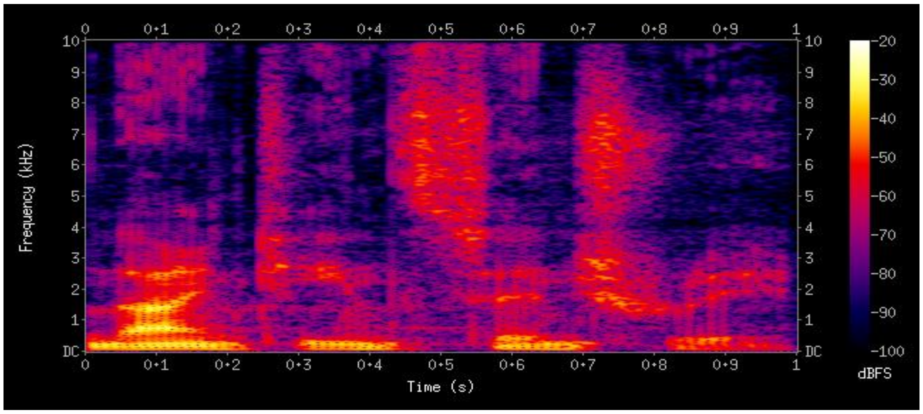


Fig. 2.5 Spectrogram of a speech signal [21].

In this case, a narrow-band spectrogram is generated. On the other hand, a short duration window gives a poor frequency resolution, but a good time resolution. This is the case, where the spectrogram is referred to as wideband spectrogram. Spectrogram contains distinctive patterns that hold different features of speech, since pure speech contains a succession of tonal (voiced) and noise-like (unvoiced) sounds. The energy is usually concentrated within the lower frequency bands for the voiced speech signal, while concentrated at the higher frequency bands for the unvoiced speech signal. The energy distribution of the speech signal in the time-frequency domain is not continuous in both spectral (vertical) and temporal (horizontal) axes. The detail energy distribution is captured as rich texture information in the spectrogram.

2.3.2 Pattern Matching Algorithms

Pattern matching is the second stage in the ASR system that performs the learning process. These algorithms determine symmetries in enormous amounts of data and helps to categorize the data into several classes. Pattern matching algorithms estimate matching scores, which reflect the degree of similarity between the input feature vectors and some stored models. The extracted features from the speech signals are used to build a model for every speaker to be stored in the database. Later, to verify a user, the matching algorithm compares the stored models with the model of the claimed speaker [24].

As mentioned above, the ASR system has two modes: training mode and testing mode. In training mode, the speaker's identity is known. The known speaker produces a phrase with a certain length that is the same for all the speakers. The feature vectors are obtained from all speakers. Then, the pattern matching algorithm labels each speaker to a certain model. In the testing phase, the speech signal input is a phrase produced by a priori unidentified speaker, and features are obtained in a similar way as in the training mode. The sequence obtained from the feature extraction step is associated with a given set of speaker models using a pattern matching algorithm. The matching algorithm produces a likelihood score for each speaker. By these statistics, a decision-based algorithm will give an estimate of the speaker's identity who produced the speech signal.

The main tasks of pattern matching are re-estimating speaker models in the training stage and producing a likelihood score for each speaker in the testing stage. There are several classification methods that can be used for pattern matching such as Vector Quantization (VQ), k-means algorithm, Hidden Markov Model (HMM), Gaussian Mixer Model (GMM), …etc. [25]. In this dissertation, two deep learning models are used for the pattern matching process.

2.4 Deep Learning Technology

The DL has earned frequent competitions in pattern recognition and Machine Learning (ML). The ML has been developed in all sectors as a technique of making machines intelligent. Simply, ML is a set of algorithms that analyze data, learn from them, and then apply what they have learned to make intelligent decisions. The DL is a sub-class of ML techniques. It is an expansion of the Neural Network (NN). The basis of DL is that, it is a multi-layer NN that contains two or more hidden layers. This makes it suitable for demonstrating very composite and greatly nonlinear relations between inputs and outputs. The DL has considerably enhanced the identification process in different areas such as speech recognition, object recognition and detection, and several other domains such as drug discovery, and genomics [26].

The DL has different models and shapes according to the application. The common network models are Deep Convolutional Neural Network (CNN) and Deep Recurrent Neural Network (RNN). The deep CNN has been widely applied in image, video, speech, and audio processing, while the RNN is used for tasks that include sequential inputs, such as speech and language [26].

2.4.1 Principles of Deep Learning

The DL is typically implemented using the NN structural design. The NN can be defined as an information processing system designed to simulate the human brain structure and functions. The word “deep” indicates the number of layers in the network as the more the layers are, the deeper the network. A customary NN has only two or three layers, while deep networks can have hundreds of layers. This allows modeling of complex data with less and expressive features. The DL is particularly right-suited in recognition applications such as voice recognition, face recognition, text translation, and innovative driver aiding systems that include route classification, and traffic sign recognition [26].

Similar to the NN, the DL entails an input layer, numerous hidden layers, and an output layer. These layers are connected by nodes, or neurons, with each hidden layer using the output of the previous layer as its input. There are several parameters that can control the performance of the deep NN. These parameters are the batch size, the number of epochs, and the number of classes besides the NN design. The network design is contingent on its width and depth which are specified by the number of layers, the number of nodes in each layer, and the activation functions in each layer. Fig 2.6 shows the basic structure of the neuron [26].

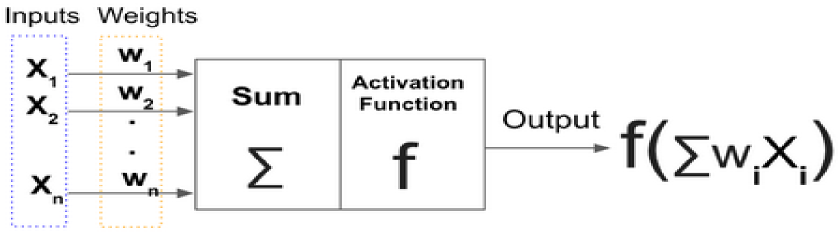


Fig 2.6 Neuron structure [27].

The NN behaves like the human brain. Biological neurons are the core components of the human brain. The main element in the NN is the neuron, which consists of a cell body, dendrites, and an axon. Each neuron receives input signals from its dendrites and produces output signals along its axon. The axon branches out and connects via synapses to the dendrites of other neurons. The neurons work, when the dendrites carry the signals to the target neuron body, where they get summed. If the final sum is above a certain threshold, the neuron gets fired, sending a spike along its axon [27].

The feed-forward NN is the simplest form of the NN as shown in Fig. 2.7. This network has four layers: an input layer, two hidden layers, and an output layer. The fully-connected characteristic means that each node is connected to all the nodes in the next layer. The number of hidden layers and their sizes are the only free parameters. The larger and deeper the hidden layers are, the more the complex patterns that can be modeled [27, 28].

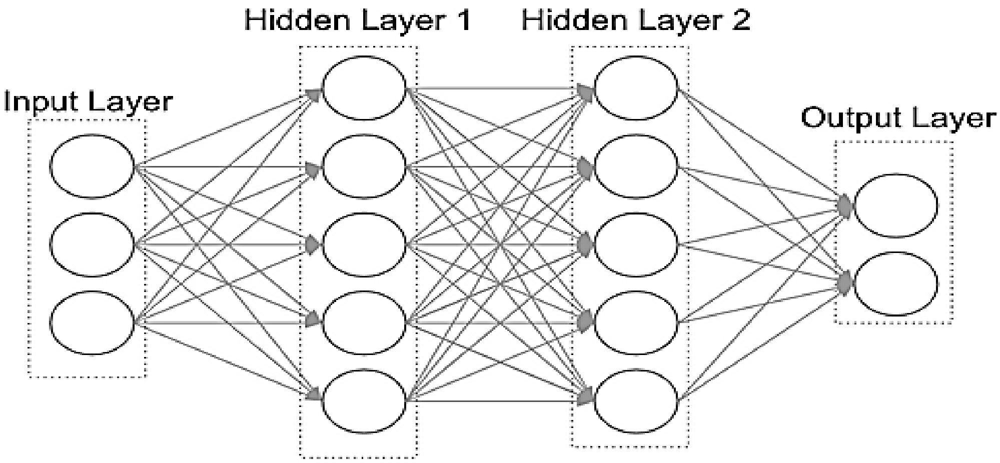


Fig 2.7 Feed-forward NN with two hidden layers [27].

2.4.2 Why Deep Learning?

Deep leaning can deal with large-size data. Adding extra layers to the conventional NN has many limitations as the network is not trained properly. So, it has poor performance. There are three main problems, undergoing the back-propagation algorithm during the training of the network; vanishing gradient, overfitting, and computational load [29].

2.4.2.1 Vanishing Gradient

The vanishing gradient occurs during the training of the NN with gradient-based learning methods and back-propagation. In such methods, each of the NN weights receives an update proportional to the gradient of the error function with respect to the current weight of each iteration in training. The problem is that, in some situations, the gradient will be vanishingly small, effectively preventing the weights from changing their values. In the worst case, this may completely stop the network from further training. In other words, the back-propagation algorithm used to train the NN propagates the output error backward to the hidden layers. The error barely extents the first hidden layer, so the weight cannot be adjusted. Therefore, the hidden layers that are near the input layer are not right trained. So, the addition of hidden layers is not useful if they cannot be trained well. Vanishing occurs, when the output error fails to reach the farther nodes [29].

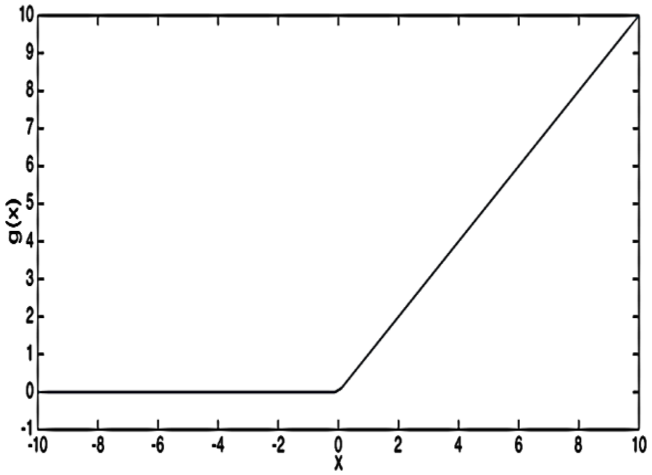


Fig 2.8 ReLU function [29]

The Rectified Linear Unit (ReLU) is an activation function used to solve the vanishing gradient. It improves the learning performance of the deep NN, it is defined as [29]:

$$\emptyset(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{2.7}$$

$$\emptyset(x) = \max(0, x) \tag{2.8}$$

2.4.2.2 Overfitting

The deep NN includes extra hidden layers as in Fig 2.9, and hence more weights. So, it becomes more complicated and susceptible to overfitting. Overfitting occurs, when the training error is larger than the testing error. Dropout can be used to avoid over-fitting, when using limited training data. The term "dropout" can be defined as dropping out units in the NN, that is temporarily removing them from the network along with all incoming and outgoing connections [29].

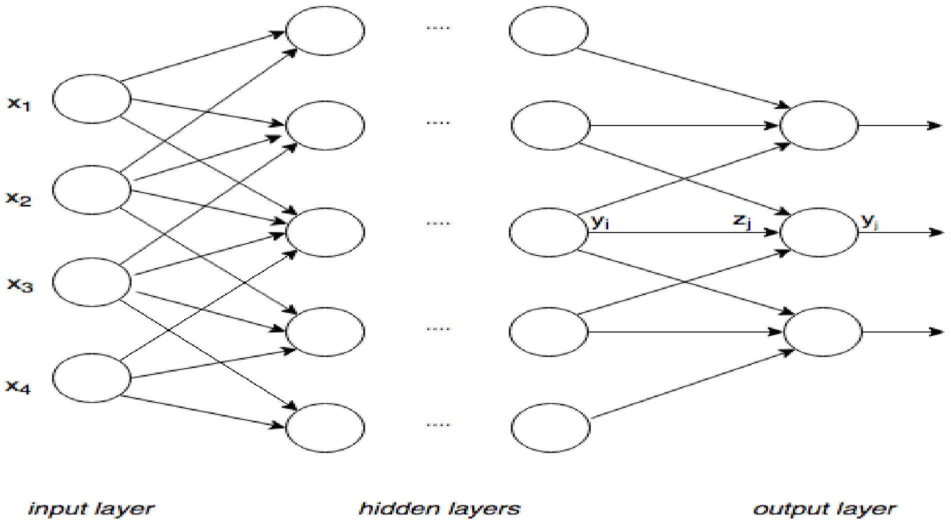


Fig. 2.9 Multi-layer neural network [29].

2.4.2.3 Computational Load

The third challenge is the time necessary to complete the training. The number of weights increases geometrically with the number of hidden layers, thus necessitating more training data. This finally entails more calculations to be made. The more computations the NN achieves, the longer the time taken for training. This is a severe problem in the real-world development of the NN [29].

2.4.3 Deep Learning Models

The DL has different models based on the used application. The two commonly used models are the RNN and CNN. The RNN has been effectively applied to various sequence forecasting and sequence classification tasks. The CNN is a good image processing tool that can be used for the visual and aural signals. Two models are used in this dissertation; the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) and the CNN [30].

2.4.3.1 Long Short-Term Memory Recurrent Neural Network (LSTM RNN)

The RNN is an NN that captures dynamic information in sequential data through cyclic connections of hidden layer nodes. It is used to classify sequential data. In many domains supplementary powerful sequential learning tools are necessary. Deep RNN has a wide use in speech processing for its ability to label sequences, which means that each input sequence is assigned to a certain class. The RNN is a NN, but with cyclic connections between its nodes. Unlike NNs, the RNN connections form a directed cycle between units. So, it can keep a context state and even store, learn, and express related information in the context of any length. Different from the traditional NN, it extends in space and time sequences. In other words, the hidden layers of the current and the next moment are related. The RNN is widely used in scenarios related to sequences such as videos consisting of image frames, audio consisting of clips, and sentences consisting of words. There are three tasks for the RNN; sequence to one learning, sequence to sequence learning, and sequence generation**.**

1. Sequence to one learning: It is a classification task, which labels a sequence of an input to a category for each sample. An important example is language identification from speech signals.

2. Sequence to sequence learning: It is the task of a mapping from sequence to sequence. An example of this task is machine translation, which translates text from one language to another languages.

3. Sequence generation: It is the modeling of the input distribution, which then can be used for synthesizing new data.

The LSTM-RNN is another version of the RNN, which replaces the hidden nodes by blocks. These blocks act as memory, so, they are called memory blocks. The main constituents of the LSTM-RNN network are a sequence input layer, an

LSTM layer, a fully-connected layer, and a softmax layer. A sequence input layer feedbacks sequence or time-series data into the network and the LSTM layers learn long-term reliance amongst time steps of sequence data. Figure 2.10 demonstrates the structure of a simple LSTM network used for the classification task. The network begins with a sequence input layer, followed by an LSTM layer. The network ends with a fully associated layer and a softmax layer to forecast set labels [31].

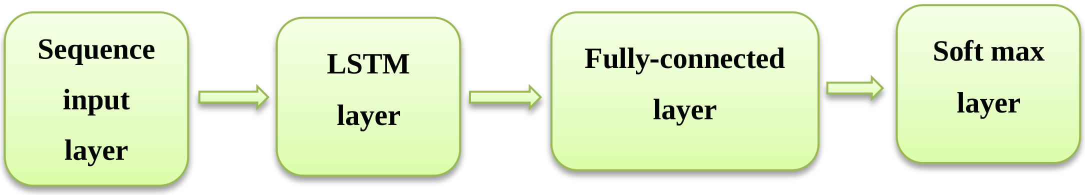


Fig. 2.10 LSTM classification network

In every layer, the hidden nodes are replaced with blocks, and every block consists of four components: memory cell (*c*), input gate (*i*), forget gate (*f*), and output gate (*o*). The cell is the memory part of the LSTM unit. The three gates control the state of the cell and the hidden state of the layer as shown in Fig. 2.11. At each time step, the layer adds information to or removes information from the cell state, where the layer controls these updates using gates. The input gate controls the update of the cell state, while the forget gate controls the reset of the cell state, and the output gate controls the flow of the output from the cell to another hidden layer [32, 33].

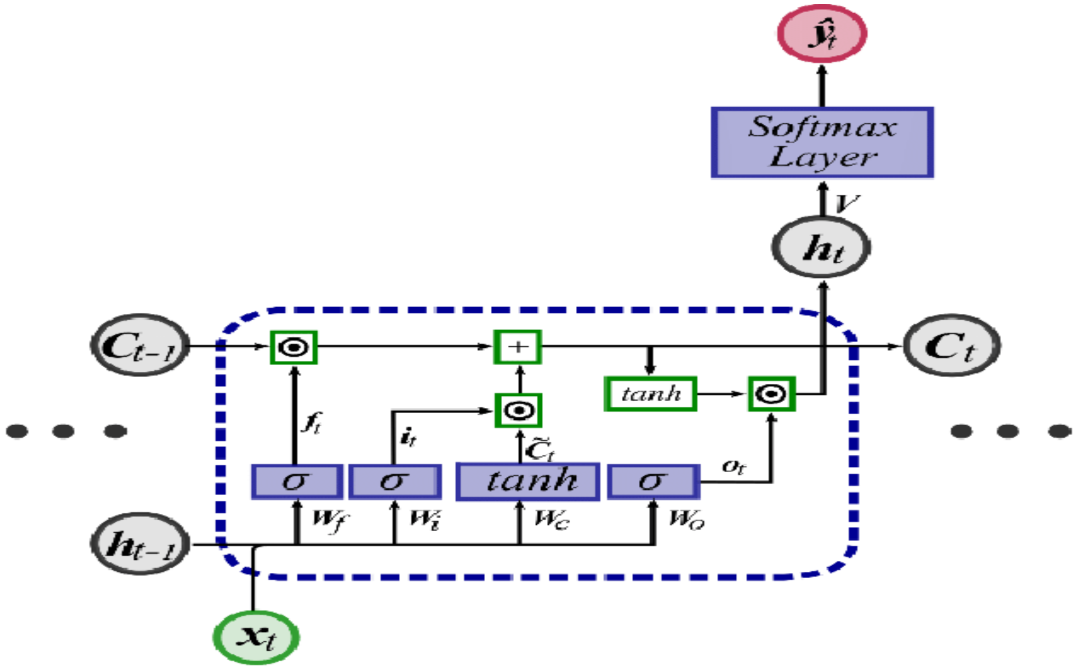


Fig. 2.11 LSTM memory block [31].

The mathematical model of the above architecture is illustrated by Equations from 2.9 to 2.16. The components, f, c, i and o are calculated as shown in the following equations [32, 33].

$$\sigma = \frac{1}{1+e^{-x}} \tag{2.9}$$

$$i_t = \sigma(\mathbf{w}_i x_t + \mathbf{R}_i h_{t-1} + b_i) \tag{2.10}$$

$$f_t = \sigma\left(\mathbf{w}_f x_t + \mathbf{R}_f h_{t-1} + b_f\right) \tag{2.11}$$

$$\acute{c}_t = tanh\left(\mathbf{w}_g x_t + \mathbf{R}_g h_{t-1} + b_g\right) \tag{2.12}$$

$$o_t = \sigma(\mathbf{w}_o x_t + \mathbf{R}_o h_{t-1} + b_o) \tag{2.13}$$

The output for the current unit is calculated by the following equation;

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2.14}$$

$$h_t = o_t \odot tanh(c_t) \tag{2.15}$$

where x_t is the input, h_{t-1} is the previous cell output, c_{t-1} is the previous cell memory, h_t is the current cell output, c_t is the current cell memory, $\mathbf{w}$, $\mathbf{R}$ are input and recurrent weights, b is the bias, $\odot$ denotes the multiplication process of vectors and σ is the sigmoid activation function.

The above equations can be simply written in the form of Eq. 2.16.

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} \mathbf{w} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b \tag{2.16}$$

Figure 2.12 illustrates the memory block. There is an input gateway on entry to determine if the entry is valid. The gate opens to enter the memory cell. The forget gate is a gateway connected to the memory gateway, and the role here is to determine whether the memory will be set to 0.

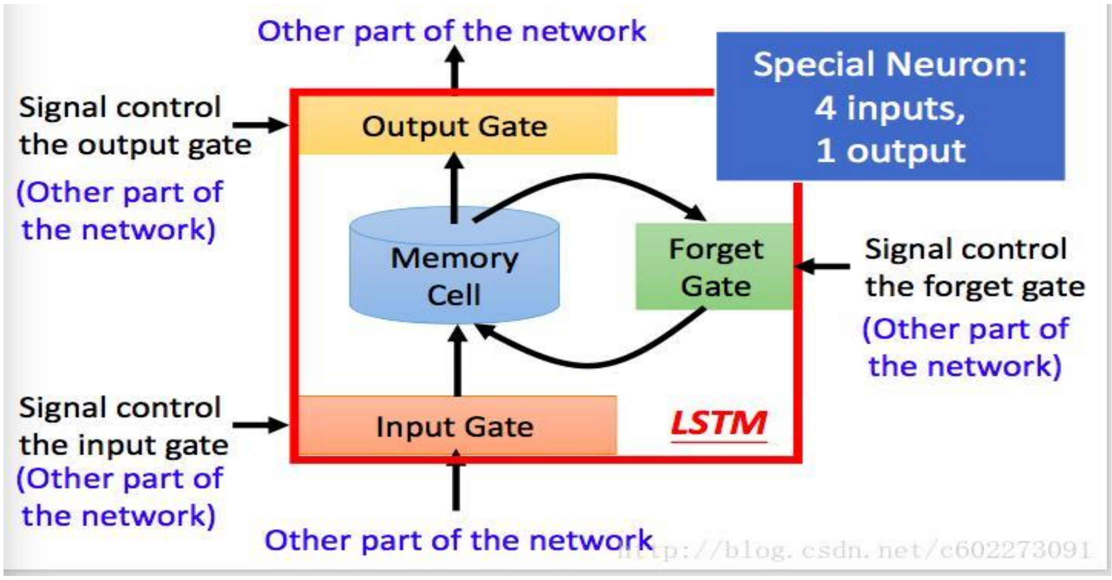

Other part of the network
Special Neuron:
4 inputs,
1 output
Signal control the output gate
(Other part of the network)
Output Gate
Memory Cell
Forget Gate
Signal control the forget gate
(Other part of the network)
Signal control the input gate
(Other part of the network)
Input Gate
LSTM
Other part of the network
http://blog.csdn.net/c602273091

Fig. 2.12 Illustration of the LSTM memory block [33]

The memory blocks in the hidden layers of the LSTM turn as a memory that reserves the current state of the network. The output of the softmax for a certain frame is a probability referring to one of the speakers as an output. It is based not only on the input frame, but also on every pre-existent frame in this sequence. The system can decide every output, dependent on the previous and present inputs [31-34].

2.4.3.2 Convolutional Neural Network

A CNN is another DL model used here for both feature extraction and pattern matching. The primary concern of CNN is image processing, since the two main operations; convolution and pooling are practically accomplished in a two-dimensional plane. So, it is instinctive to organize the input as a two-dimensional (2-D) array, being the pixel values at the (horizontal and vertical) coordinates indices, this is a term taken from image processing applications. The reality that the CNN turns the manual feature extraction design into the automated process is its primary feature and advantage. The CNN comprises the feature extractor in the training process rather than manipulating it manually. The CNN applies a small window over the input image at both training and testing times, so that the weights of the network that look through this window can learn from various features of the input data irrespective of their absolute position within the input. The feature extractor of the CNN comprises special types of NNs, at which the weights are settled by the training process [35].

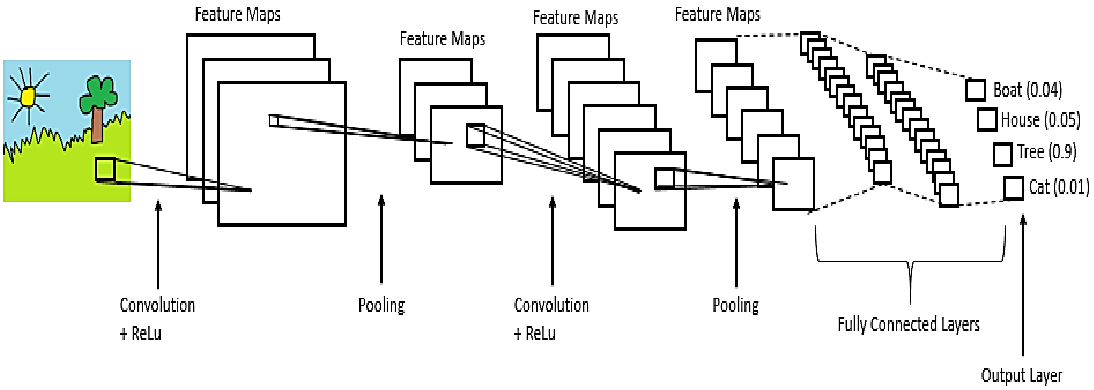
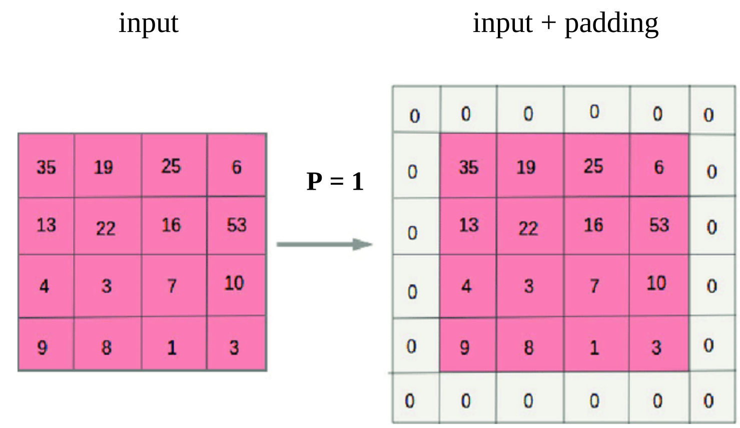


Fig. 2.13 CNN structure [35].

The CNN involves two types of NN: one for features extraction from the input image and another to classify these features. The feature extraction NN consists of pairs of a convolutional layer and a pooling layer. The convolutional layer, as implied in its name, processes the image using the convolution operation. It can be thought of as a collection of digital filters. The pooling layer combines the neighboring pixels into a single pixel, and therefore it reduces the dimensions of the image. These are the main distinctions between the CNN and other NNs. In summary, the CNN consists of a serial connection of the feature extraction network and the classification network. The convolutional layer converts the images via the convolution operation, and the pooling layer reduces the dimensions of the images. The classification network usually employs the ordinary multi-class classification NN. The main component of the CNN is explained in detail in the following section.

2.4.3.2.1 Convolutional Layer

Convolutional Layer referred to as Conv. layer represents the starting point of the CNN and carries out the main operations of training, and thus firing the neurons of the network. Before defining the convolution operation, two basic operations are defined: padding and stride. Padding means adding extra pixels around the image in order to take the pixels on the edges into consideration. As, the pixels on the corner of the image are less used than the pixels in the middle of the image, which means that the information from the edges is ignored when the convolution product is performed. A padding with zeros is often used. The following figure (Fig. 2.14) illustrates the padding of a 2D matrix with $p = 1$, where p is the number of added pixels on each of the four sides of the image [36].



35	19	25	6
13	22	16	53
4	3	7	10
9	8	1	3

0	0	0	0	0	0
0	35	19	25	6	0
0	13	22	16	53	0
0	4	3	7	10	0
0	9	8	1	3	0
0	0	0	0	0	0

Fig. 2.14 Padding with zeros [37].

The second operation is stride, which is defined as the step taken in the convolution product. A large stride allows to shrink the size of the output and vice versa. Since the stride is the number of pixel shifts over the input matrix. A filter with $s = 1$ means that the filter moves one pixel during the convolution operation, where s is the stride parameter.

Convolutional filter accomplishes the convolution operation over the input dimensions and the resultant image is called the feature map. The convolution product on a 2D matrix is the sum of the element-wise product that consists of a three dimensional arrangement of neurons (a stack of 2-dimensional layers of neurons, one for each channel depth) as shown in Fig. 2.15 [37]. This can be mathematically represented by Eq. 2.17. For an image $I(m, n, c)$ and kernel $K(f, f, c)$, the convolutional product between the image and the filter is [37]:

$$conv(I, K)_{x,y} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{c} K_{i,j,k}\, I_{x+i-1, y+j-1, k} \qquad (2.17)$$

The dimensions of the output image after convolution will be [37]:

$$dim\big(conv(I, K)\big) = \left(\left\lfloor \frac{(m+2p-f)}{s} + 1 \right\rfloor, \left\lfloor \frac{(n+2p-f)}{s} + 1 \right\rfloor, c\right) \ ; s > 0 \qquad (2.18)$$

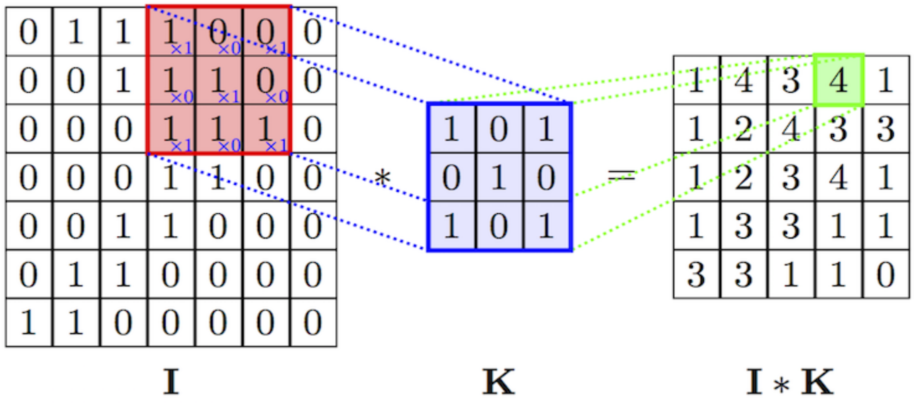


Fig. 2.15 Illustration of the convolution process [37].

where $\lfloor \quad \rfloor$ is the floor function and m, f, p, and s are the image size, kernel size, padding size, and number of stride, respectively.

The convolutional layer creates new images called feature maps. The feature map highlights the unique features of the original image. In brief, the convolutional layer performs the convolutional filters on the input image and yields the feature maps. The extracted features in the convolutional layer are determined by the trained convolutional filters. Consequently, the features extracted by the convolutional layer differ based on the used convolutional filter. The feature maps created by the convolutional filter are handled over the activation function before the layer yields the output. The activation function of the convolutional layer is similar to that of the conventional neural network. However, in the majority of the current applications, the ReLU function is used, the sigmoid function and the tanh function are frequently used as well [36, 37], in this thesis a ReLU activation function is used. The output after activation is given by:

$$Output = \emptyset\left(conv(I,K)_{x,y}\right) = \emptyset\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} K_{i,j,k}\, I_{x+i-1,y+j-1,k} + b\right) \qquad (2.19)$$

where $\emptyset$ is the activation function and b is the bias. The bias is an additional parameter in the network, which is used to adjust the output along with the weighted sum of inputs to the neuron. Therefore, the bias is constant, which helps the model in a way that can fit best for the given data.

2.4.3.2.2 Pooling Layer

The pooling layer is typically placed after the convolutional layer. Its main function lies in decreasing the spatial dimensions (Width × Height) of the input for the next convolutional layer. The process performed by this layer is also known as 'down-sampling', as the reduction of the size leads to loss of information the pooling process is described in Fig. 2.16 [36, 37].

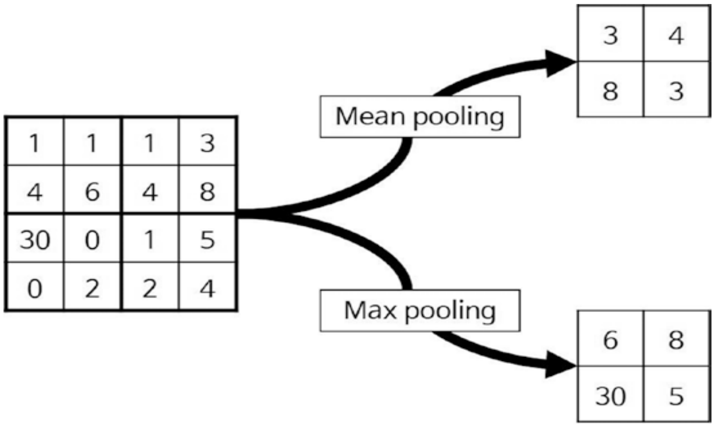


Fig. 2.16 Pooling with two different methods [37].

The output size after pooling is:

$$dim\big(pooling(image)\big) = \left(\left\lfloor \frac{(m+2p-f)}{s} + 1 \right\rfloor , \left\lfloor \frac{(n+2p-f)}{s} + 1 \right\rfloor , c \right) \quad ; s > 0 \qquad (2.20)$$

Nevertheless, such a loss is advantageous for the network for two motives: the reduction in size leads to less computational overhead for the forthcoming layers of the network, and also it works against overfitting. Much similar to the convolution task done above, the pooling layer takes a sliding window or a certain area that is moved in stride across the input converting the values into demonstrative values. The transformation is either implemented by taking the maximum value from the values observed in the window (called max pooling), or by taking the average of the values (called mean pooling). Max pooling is preferred over other pooling types due to its good performance [37].

As mentioned above, the pooling layer comprises both the activation operation that is applied to each element after convolution and subsampling of the data after activation. The idea of using the activation function is to increase the non-linearity in our images since it is a non-linear operation. Activation operation is typically a nonlinear operation that is applied to each element of convolution output such as $max(0, x)$ which is also called ReLU or $1 / (1 + e^{x})$, where x is the input data. The activation operation does not change the size of the input. Subsampling is applied after activation that reduces the size of the input to typically $1/2$ at each dimension. Window size that is used during subsampling is 2×2 or 3×3. If the input data size is (W, H) then the size after pooling will be $(W / 2, H / 2)$ if subsampling by $1/2$ is used [38].

2.4.3.2.3 Fully-Connected Layer:

The Fully-Connected (FC) layer reduces the size of the input data to the size of classes that the CNN is trained for by combining the output of the convolutional layer with different weights. Each neuron at the output of the convolutional layer will be connected to all other neurons after being weighted properly. Similar to the convolutional layer, the weights of these taps in the FC layer are found through the back-propagation algorithm. An optimization method is used to adjust the parameters in order to reduce the cost function. The Stochastic Gradient Decent (SGD) method is used for optimization. Cross entropy is used as a loss or a cost function that evaluates the distances between the real and predicted values on a single point [37, 38].

2.4.3.2.4 Classification Layer (CL):

The CL it is the final layer of the CNN that converts the output of the FC to a probability of each object being in a certain class. Typically, the softmax algorithm is used in this layer, which transforms the numerical results of the network to a probability. The softmax maps output results to the interval (0, 1), in which sum of all outputs is equal one [36-38]. The CNN is similar to the NN, where neurons are used in the FC layer, but they are replaced by the convolution operation at the initial layers. Such a replacement is performed to focus on a similar features at different positions in the image, which is achieved by convolutional filters.

2.4.3.3 Benchmark Model

The Benchmark model is a simple single-layer CNN model. This model consists of one convolutional layer with 32 kernel filters followed by one max pooling layer. In this dissertation, our proposed models will be compared with the benchmark model.

2.5 Related Work to Speaker Recognition

There are many algorithms and feature extraction methods that have been used in the ASR system. Robert Togneri and Daniel pullella [2011] have implemented a text-independent ASR based on GMM and GMM-UBM as classifiers with MFCC features. They test their system in two scenarios: clean speech and noisy speech at different SNR values. The system accuracy reaches 94.5% for clean speech and 74.2 for noisy speech at 30 dB [19].

Yaming Wang [2012] has proposed an ASR system that is based on MFCC features with Vector Quantization (VQ) modeling technique. It integrates a hearing masking effect and a group of dozen triflers into a traditional MFCC feature extraction for robust speaker identification. The masker can decrease the influence of noise on the speech signal, and improve the recognition rate. The mixture of triflers can enhance high-frequency calculation accuracy. The results show that the improved algorithm is able to effectively overcome the environment noise and the variation of speaker’s voice with a higher recognition rate of voice signal to some extent [39].

The DL has been also used for ASR systems in two main strategy either for feature extraction or classification. Y. Lukic et al., [2016] use the spectrogram as an input to the CNN, and then they studied the design of the network for identification and clustering of 10 speakers. The achieved accuracy was 97% [40].

Nayana P.K. et al. [2017] have implemented a text-independent ASR system using GMM and i-vector method with two features: Power Normalized Cepstral Coefficients (PNCCs) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP) coefficients. The GMM is a parametric model that provides the Probability Density Function (PDF) of the speaker model. The Gaussian model can be described by three parameters: mean, weight, and covariance [41].

$$f(x_n/\lambda) = \sum_{g=1}^{M}(\pi_g N(x_n/\mu_g, \Sigma_g) \tag{2.21}$$

where π_g, μ_g and Σ_g indicate the weight, mean vector, and covariance matrix. For a set of features $(X = \{xn|n \in 1 \ldots .T\})$. The probability of observing these features is:

$$p(X|Y) = \prod_{n=1}^{T} P(x_n|\lambda) \tag{2.22}$$

where $\lambda = \{\pi_g, \mu_g, \Sigma_g\},\ g = 1, \ldots, M$

Since the PDF for a specific speaker is provided, the probability score can be obtained. A Universal Background Model (UBM) is created from speech signals of different speakers. The test model is compared against all the speaker models and the Log Likelihood Ratio (LLR) is calculated. The model with the highest LLR is the test model [41].

$$LLR = \sum_{t=1}^{T}(\log\{p(x_t/\lambda_{ref})\} - \log\{p(x_t/\lambda_{UBM})\}) \tag{2.23}$$

The i-vector method is derived from the GMM super vector. It is implemented using two types of classifiers: Cosine Distance Scoring (CDS) and

Probabilistic Linear Discriminant Analysis (PLDA). The PLDA has better classification accuracy than the CDS, and also the PNCC has better identification of speakers even for noisy speech [41].

Yanpei Shi et al. [2020] propose a two stage attention model (time attention and frequency attention) used with Time Delay Neural Networks (TDNNs) and Convolutional Neural Networks (CNNs). They study the effect of noise on their proposed model. For the TDNN based models, the recognition rate reaches 91.1%, while reaches 92.0% for the CNN based architecture. Also the frequency attention model in both TDNN and CNN is better than the time attention model [42].

CHAPTER 3

MATHEMATICAL MODELS

3.1 Introduction

Most studies in ASR applications consider an ideal scenario for training and testing without noise, reverberation or interference. In real cases, some sources of degradation may exist. So, in this dissertation, we consider text-independent speaker recognition in the realistic case of degradation. The effects of noise, reverberation, and interference are considered in this dissertations. Moreover, speech enhancement techniques such as spectral subtraction and wavelet denoising are considered as a pre-processing stage, to enhance the performance of the speaker recognition process. In addition, Radon Transform (RT) is used for better representation of speech signals in the presence of noise as it is robust to the noise effect. Also, the interference effect is cancelled with a signal separation algorithm. This chapter presents all mathematical analysis and models for all used algorithms and methods for these purposes in this thesis.

3.2 Spectral Subtraction

Spectral subtraction is one of the mostly used speech enhancement method. The clean speech signal spectrum can be obtained by subtracting the estimated noise spectrum from noisy speech spectrum [43, 44]. Consider $s(n)$ as the clean speech, $y(n)$ as the noisy speech, and $v(n)$ as the noise.

$$y(n) = s(n) + v(n) \tag{3.1}$$

Applying Fast Fourier Transform (FFT) on Eq. 3.1 yields [45].

$$Y(\omega) = S(\omega) + V(\omega) \tag{3.2}$$

As $Y(\omega) = |Y(\omega)|e^{j\theta_y}$ and $V(\omega) = |V(\omega)|e^{j\theta_v}$

Hence, the estimated spectrum of the speech signal can be represented as:

$$\hat{S}(\omega) = \{|Y(\omega)| - |V(\omega)|\}e^{j\theta_y(\omega)} \tag{3.3}$$

where $\hat{S}(\omega)$ is the estimated spectrum of the clean speech signal, and $V(\omega)$ is estimated by taking the average value during non-speech periods of the signal. After taking the Inverse Fast Fourier Transform (IFFT), an estimate of the clean signal can be obtained. This algorithm has low complexity [45].

3.3 Wavelet Denoising (Wden)

Wavelet denoising is another enhancement technique, which depends on using Discrete Wavelet Transform (DWT) to decompose the signal into approximation and detail coefficients. Then, noise is removed by thresholding of the detail coefficients, and finally performing the inverse wavelet transform to restore the clean signal [46]. There are two thresholding methods: soft and hard thresholding.

For hard thresholding, the equation is [46]:

$$f_{hard}(x) = \begin{cases} x & |x| \geq T \\ 0 & |x| < T \end{cases} \tag{3.4}$$

For soft thresholding, the equation is:

$$f_{soft}(x) = \begin{cases} x & |x| \geq T \\ 2x - T & T/2 \leq x < T \\ T + 2x & -T < x \leq -T/2 \\ 0 & |x| < T/2 \end{cases} \tag{3.5}$$

where T is the threshold value and x is the detail coefficients of the DWT. In this thesis soft thresholding is used.

3.4 Radon Transform

Radon transform (RT) is a good tool that picks up the directional features of a 2-D signal. The spectrogram of the speech signal is obtained as an image representation of the power spectrum. This image is processed with the Radon transform. The spectrogram includes acoustic features such as energy, pitch, fundamental frequency, formants and time in the form of a pattern. The RT reveals the variations in the image due to the change in energy, pitch, fundamental frequency, formants, etc. The RT is able to capture these features in the pattern by projecting it onto different orientation slices. The Radon projection is implemented by adding all intensity values of the pixels in the spectrogram images inside the circle that encompasses the pattern to be recognized. The Radon projection is computed by adding the pixel intensity values in the spectrogram image along a certain direction at a specific displacement [47].

Given a 2-D image, $f(x,y)$ the RT is obtained as follows:

$$R(r,\theta) = \Re[f(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(r - xcos\theta - ysin\theta)dx\, dy. \qquad (3.6)$$

where r is the distance of a line from the origin, $\theta \in [0,\pi]$ is the angle between the distance vector and x-axis and $\delta(.)$ is the Dirac function. The symbol $\Re$ represents the RT operator. The RT computes the Radon projections of the spectrogram in different directions. In each projection, the variations of the pixel intensities are preserved evenly, though the pixels are far from the origin. The main characteristic of using the RT is its insensitivity to noise and reverberation.

If we consider that the noise corrupting an image $f(x,y)$ is $\eta(x,y)$ as shown in the following equation [48]:

$$\hat{f}(x,y) = f(x,y) + \eta(x,y) \qquad (3.7)$$

where $\eta(x,y)$ is zero-mean white noise. The RT of $\hat{f}(x,y)$ is given by:

$$\Re[\hat{f}(x,y)] = \Re[f(x,y)] + \Re[\eta(x,y)] \qquad (3.8)$$

Since the RT is line integrals of the image, for the continuous case, the Radon transform of white noise is constant for all points and directions, and it is equal to the mean value of the noise (if integrated over an infinite axis), which is assumed to be zero. In this case, the signal spectrogram (digital image) is discontinuous and it is composed of a finite number of pixels. Hence, $\Re[\eta(x,y)]$ does not become zero. Radon projection angles change from 0 to 180. The angle is taken here from 15° to 180° with a constant interval of 15. After performing Radon projection, a 2-D DCT is performed to reduce feature dimensions. It compacts the energy of the signal in the first few coefficients. The 2-D DCT of an image $f(m,n)$ is given as [47-49]:

$$F(M,N) = \propto_p \propto_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) \cos\frac{\pi(2m+1)p}{2M} \; cos\frac{\pi(2n+1)q}{2N} \qquad (3.9)$$

$$0 \le p \le M-1, \qquad 0 \le q \le N-1$$

$$\propto_p = \frac{1}{\sqrt{M}}, \quad P = 0 = \sqrt{\frac{2}{M}}, \quad 1 \le p \le M-1,$$

$$\propto_q = \frac{1}{\sqrt{N}}, \quad q = 0 = \sqrt{\frac{2}{N}}, \quad 1 \le q \le N-1$$

where $\propto_p$ and $\propto_q$ are the normalization factors; p and q are the frequencies; M and N are the numbers of rows and columns of $f(m,n)$.

3.5 Signal Separation Algorithm

Blind signal separation is an essential branch of signal processing. It deals mostly with mixed signals which are often encountered in real life. Real-life speech signals are frequently mixed with undesired signals. This fact has motivated the evolution of blind signal separation algorithms. The word blind means that there is no a priori information about the mixed signals and their sources. The blind signal separation problem or blind source separation has been driven by practical problems represented in multiple sources and multiple sensors. They have a common objective, which is to separate and estimate the source signals without knowledge of the characteristics of the transmission channel [50, 51]. The problem is illustrated in Fig. 3.1.

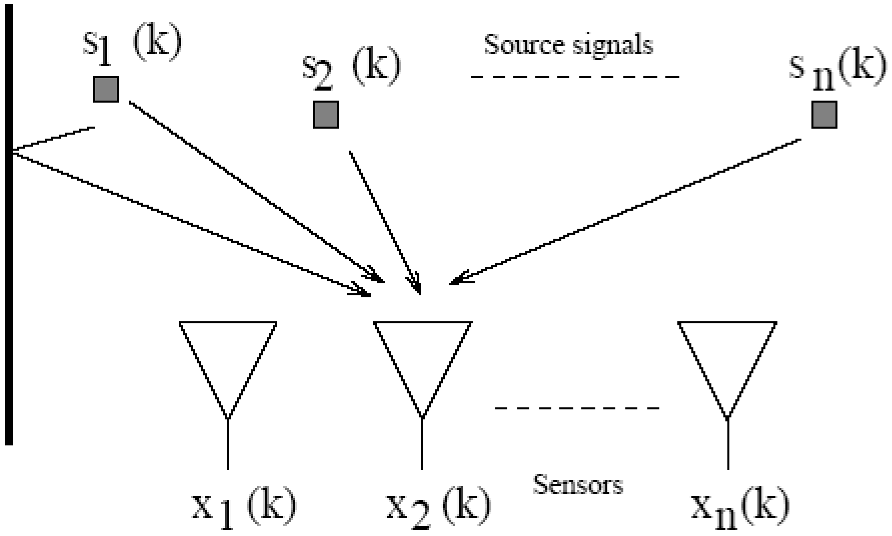


Fig. 3.1 A typical multi-source and multi-sensor situation

The fundamental and challenging problem is the separation of independent sources from mixed observed data. It has wide applications such as hearing aids, noise cancellation, voice controlled devices, speaker identifiers and speech enhancement. Blind signal separation deals with mixtures of signals [52]. The focus is the separating two sources from two mixtures (2×2 system). This is the simplest case of the general multi-source multi-input $n \times n$ problem. The separation method is based on the use of output decorrelation as the signal separation criterion. If there are two signal sources $s_1(k)$, $s_2(k)$ and two observations $x_1(k)$, $x_2(k)$ in a 2×2 LTI system, the source signals are assumed to be zero mean and statistically independent. The observations are assumed to be convolutive sums of the sources as shown in Fig 3.2 and represented by the following equations [50-53]:

$$x_1(k) = \sum_{i=0}^{p} h_{11}(i) s_1(k-i) + \sum_{i=0}^{p} h_{12}(i) s_2(k-i) \tag{3.10}$$

$$x_2(k) = \sum_{i=0}^{p} h_{21}(i) s_1(k-i) + \sum_{i=0}^{p} h_{22}(i) s_2(k-i) \tag{3.11}$$

Or in vector-matrix form as follows:

$$\begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{11}^{\mathrm{T}} & \mathbf{h}_{12}^{\mathrm{T}} \\ \mathbf{h}_{21}^{\mathrm{T}} & \mathbf{h}_{22}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} s_1(k) \\ s_2(k) \end{bmatrix} \tag{3.12}$$

where $\mathbf{h}_{ij}^{T} = \left[h_{ij}(0), \ldots\ldots h_{ij}(p)\right]$ and $s_i^T(k) = [s_i(k), \ldots\ldots s_i(k-p)]$

$\mathbf{h}_{ij}$ is the impulse response from source j to sensor i and p is the order of the filter. For simplicity, the source signals are assumed to be zero-mean and statistically-independent. From Eq. 3.10 and 3.11, it is clear that the mixtures are convolutive sums of sources in the presence of noise. The problem is simplified by assuming that the signals arrive at the sensors unfiltered, which is equivalent to setting $\mathbf{h}_{11} = \mathbf{h}_{22} = 1$.

Taking the z-transform of Eq. 3.12 yields:

$$\begin{bmatrix} X_1(z) \\ X_2(z) \end{bmatrix} = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix} \begin{bmatrix} S_1(z) \\ S_2(z) \end{bmatrix} \tag{3.13}$$

Simplifying Eq. 3.13 leads to:

$$\begin{bmatrix} X_1(z) \\ X_2(z) \end{bmatrix} = \begin{bmatrix} 1 & H'_{12}(z) \\ H'_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} S'_1(z) \\ S'_2(z) \end{bmatrix} \tag{3.14}$$

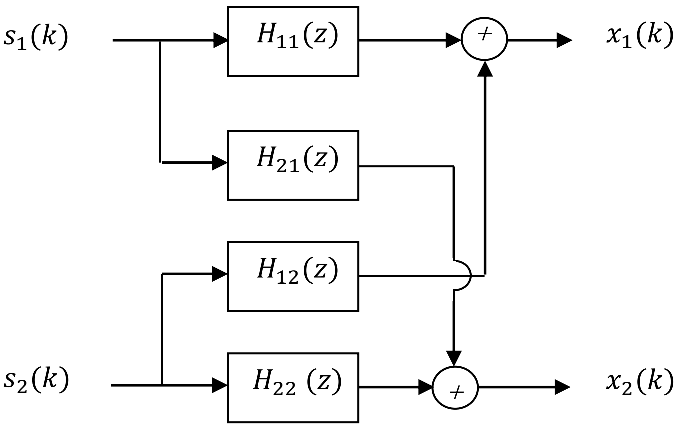


Fig. 3.2 A fully-coupled 2 × 2 mixing system [53].

where

$$S_1'(z) = H_{11}(z)\, S_1(z)$$

$$S_2'(z) = H_{22}(z)\, S_2(z)$$

$$H_{12}'(z) = \frac{H_{12}(z)}{H_{22}(z)}$$

$$H_{21}'(z) = \frac{H_{21}(z)}{H_{11}(z)}$$

where $s_1(k)$ and $s_2(k)$ are the true source signals and $\mathbf{h}_{ij}$ are the true impulse responses of sources to sensors. $\grave{s}_i(k)$ is then the signal as observed by the i_{th} sensor. It is assumed that $H_{ii}(z) = 1$, and thus $\grave{s}_i(k) = s_i(k)$, and $\grave{H}_{ij}(z) = H_{ij}(z)$.

For $H_{ii}(z) = 1$, which is the case of interest, Eq. 3.14 is simplified to:

$$\begin{bmatrix} x_1(z) \\ x_2(z) \end{bmatrix} = \begin{bmatrix} 1 & H_{12}(z) \\ H_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} S_1(z) \\ S_2(z) \end{bmatrix} \tag{3.15}$$

The objective of blind signal separation is to get the signals $y_1(k)$ and $y_2(k)$, which are as close as possible to $x_1(k)$ and $x_2(k)$. We can assume that:

$$\begin{pmatrix} y_1(k) \\ y_2(k) \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{w}_1^T \\ \mathbf{w}_2^T & 1 \end{pmatrix} \begin{pmatrix} x_1(k) \\ x_2(k) \end{pmatrix} \tag{3.16}$$

where

$$\mathbf{w}_i^T = [w_i(0), \dots \dots, w_i(q)]$$

$$\mathbf{x}_i^T(k) = [x_i(k), \dots, x_i(k-q)]$$

then

$$\begin{bmatrix} Y_1(z) \\ y_2(z) \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{W}_1(z) \\ \mathbf{W}_2(z) & 1 \end{bmatrix} \begin{bmatrix} X_1(z) \\ X_2(z) \end{bmatrix} \tag{3.17}$$

Substituting Eq. 3.15 into Eq. 3.17 leads to:

$$\begin{bmatrix} Y_1(z) \\ y_2(z) \end{bmatrix} = \begin{bmatrix} 1 + \mathbf{W}_1(z)H_{21}(z) & \mathbf{W}_1(z) + H_{12}(z) \\ \mathbf{W}_2(z) + H_{21}(z) & 1 + \mathbf{W}_2(z)H_{12}(z) \end{bmatrix} \begin{bmatrix} S_1(z) \\ S_2(z) \end{bmatrix} \tag{3.18}$$

The objective of the problem is to find a suitable $\mathbf{w}_i(z)$ such that $Y_1(z)$ and $Y_2(z)$ each contains $S_1(z)$ or $S_2(z)$ only.

3.5.1 Block-Based Separation Algorithm for a Convolutive System

This section deals with a time-domain iterative separation algorithm for the 2×2 convolutive system as shown in Fig. 3.3. The algorithm intends to minimize the output cross-correlations for an arbitrary number of lags with $q + 1$ tap Finite Impulse Response (FIR) filters [53].

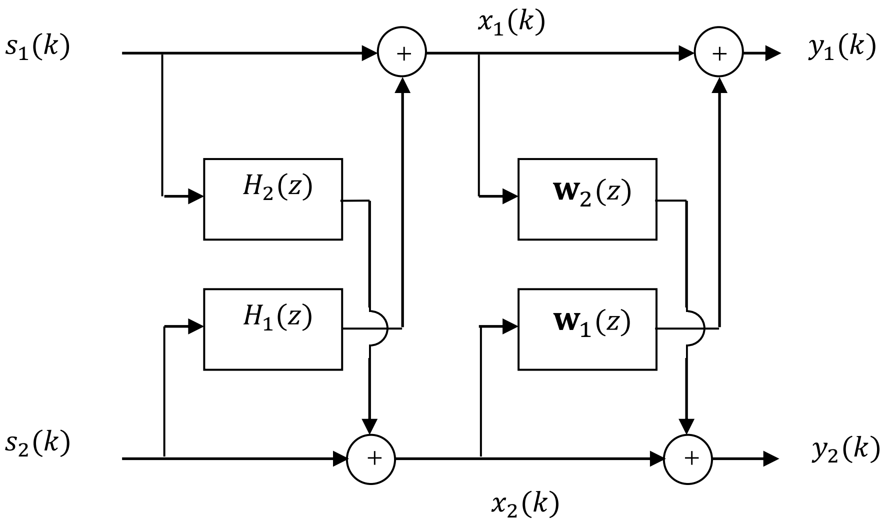


Fig. 3.3 Schematic diagram of the 2×2 separation algorithm [52].

From Eq. 3.19, we find that the solution of the problem is to find a suitable $\mathbf{w}_1(z)$ and $\mathbf{w}_2(z)$ such that $Y_1(z)$ and $Y_2(z)$ each contains only $S_1(z)$ or $S_2(z)$. This is achieved only if either the diagonal or the anti-diagonal elements are zero. Assuming $s_1(k)$ and $s_2(k)$ are stationary, zero-mean and independent random signals, the cross-correlation of the two signals is equal to zero, that is

$$r_{s_1 s_2}(l) = E[s_1(k)s_2(k+l)] = 0 \qquad \forall \quad l \tag{3.19}$$

If $y_1(k)$ and $y_1(k)$ each contains components of $s_1(k)$ or $s_2(k)$ only, then the cross-correlation of $y_1(k)$ and $y_2(k)$ should also be zero. Then

$$r_{y_1 y_2}(l) = E[y_1(k)y_2(k+l)] = 0 \qquad \forall \quad l \tag{3.20}$$

Substituting Eq. 3.16 into Eq. 3.20 gives:

$$r_{y_1 y_2}(l) = E[(x_1(k) + \mathbf{w}_1^T x_2(k))(x_2(k+l) + \mathbf{w}_2^T x_1(k+1))] \tag{3.21}$$

Denote $r_{x_i x_j}(l) = E[x_i(k)x_j(k+l)$. Eq. 3.21 becomes:

$$r_{y_1 y_2}(l) = r_{x_1 x_2}(l) + \mathbf{w}_1^T \begin{bmatrix} r_{x_2 x_2}(l) \\ \vdots \\ r_{x_2 x_2}(l+q) \end{bmatrix} + \mathbf{w}_2^T \begin{bmatrix} r_{x_1 x_1}(l) \\ \vdots \\ r_{x_1 x_1}(l+q) \end{bmatrix} + \mathbf{w}_1^T \boldsymbol{R}_{x_2 x_1}(l)\mathbf{w}_2 \tag{3.22}$$

where $\mathbf{R}_{x_2 x_1}(l) = \mathrm{E}[\mathbf{x}_2(k)\big(\mathbf{x}_1(k+l)\big)^T]$ is a $(q+1) \times (q+1)$ matrix which is a function of the cross-correlation of x_1 and x_2.

The cost function C is defined as the sum of the squares of the cross-correlations between the two inputs as [53]:

$$C = \sum_{l=l_1}^{l_2} r_{y_1 y_2}(l)^2 \tag{3.23}$$

where l_1 and l_2 are arbitrarily chosen ranges of cross-correlation lags, and C can be also expressed as:

$$C = \mathbf{r}_{y_1 y_2}^T \mathbf{r}_{y_1 y_2} \tag{3.24}$$

where

$$\boldsymbol{r}_{\boldsymbol{y_1 y_2}} = [r_{\boldsymbol{y_1 y_2}}(l_1), \dots, r_{\boldsymbol{y_1 y_2}}(l_2)]^{\boldsymbol{T}} \tag{3.25}$$

which equals:

$$\mathrm{r}_{y_1y_2} = \mathrm{r}_{x_1x_2} + \left[\mathbf{Q}^{+}_{x_2x_2}\right]^{T}\mathbf{w}_1 + \left[\mathbf{Q}^{-}_{x_1x_1}\right]^{T}\mathbf{w}_2 + \mathbf{R}^{T}_{x_2x_1}\mathbf{A}(\mathbf{w}_2)\mathbf{w}_1 \tag{3.26}$$

or

$$\mathrm{r}_{y_1y_2} = \mathrm{r}_{x_1x_2} + \left[\mathbf{Q}^{+}_{x_2x_2}\right]^{T}\mathbf{w}_1 + \left[\mathbf{Q}^{-}_{x_1x_1}\right]^{T}\mathbf{w}_2 + \mathbf{R}^{T}_{x_1x_2}\mathbf{A}(\mathbf{w}_1)\mathbf{w}_2 \tag{3.27}$$

where $\mathbf{Q}^{+}_{x_2x_2}$ and $\mathbf{Q}^{-}_{x_1x_1}$ are $(q+1)\times(l_2-l_1+1)$ matrices, $\boldsymbol{R}_{x_2x_1}$ is a $(2q+1)\times(l_2-l_1+1)$ matrix. These are all correlation matrices of x_1 and x_2 and are estimated using sample correlation estimates. $\mathbf{A}(\mathbf{w}_1)$ and $\mathbf{A}(\mathbf{w}_2)$ are $(2q+1)\times(q+1)$ matrices which contain $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively. In order to find some suitable $\mathbf{w_1}$ and $\mathbf{w}_2$, C is minimized such that:

$$\frac{\partial C}{\partial w_i} = [0,\ldots,0]^{T}, \quad i = 1,2 \tag{3.28}$$

Let:

$$\psi_1 = (\left[\mathbf{Q}^{+}_{x_2x_2}\right]^{T} + \mathbf{R}^{T}_{x_2x_1}\mathbf{A}(\mathbf{w}_2)) \tag{3.29}$$

$$\psi_2 = (\left[\mathbf{Q}^{-}_{x_1x_1}\right]^{T} + \mathbf{R}^{T}_{x_1x_2}\mathbf{A}(\mathbf{w}_1)) \tag{3.30}$$

Substituting Eq. 3.29 and 3.30 into Eq. 3.26 and 3.27 gives:

$$r_{y_1y_2} = r_{x_1x_2} + \psi_1\mathbf{w}_1 + \left[\mathbf{Q}^{-}_{x_1x_1}\right]^{T}\mathbf{w}_2 \tag{3.31}$$

or

$$r_{y_1y_2} = r_{x_1x_2} + \psi_2\mathbf{w}_2 + \left[\mathbf{Q}^{+}_{x_2x_2}\right]^{T}\mathbf{w}_1 \tag{3.32}$$

From Eq. 3.27, we obtain:

$$\mathbf{w}_1 = -(\psi_1^{T}\psi_1)^{-1}\psi_1^{T}(r_{x_1x_2} + \left[\mathbf{Q}^{-}_{x_1x_1}\right]^{T}\mathbf{w}_2) \tag{3.33}$$

$$\mathbf{w}_2 = -(\psi_2^{T}\psi_2)^{-1}\psi_2^{T}(r_{x_1x_2} + \left[\mathbf{Q}^{+}_{x_2x_2}\right]^{T}\mathbf{w}_1) \tag{3.34}$$

$\mathbf{w}_1$ and $\mathbf{w}_2$ can be obtained by iterating between the two equations until convergence is achieved, when the rate of change of parameter values is less than a pre-set threshold, e.g. 0.01% [53]. By estimating $\mathbf{w_1}$ and $\mathbf{w}_2$, we then obtain a set of outputs $y_1(k)$ and $y_2(k)$. Each output contains $s_1(k)$ or $s_2(k)$ only [53], and the flowchart is shown in Fig. 3.4.

Algorithm steps:

1- The cross correlation matrices $\mathbf{Q}^{-}_{x_1x_1}$ and $\mathbf{Q}^{+}_{x_2x_2}$ are initialized.

2- The functions ψ_1 and ψ_2 are constructed.

3- The weights $\mathbf{w}_1$ and $\mathbf{w}_2$ are updated.

4- Check if convergence is achieved.

5- The weights $\mathbf{w}_1$ and $\mathbf{w}_2$ are updated iteratively till the cost function C is minimized and convergence occurs.

6- The weights $\mathbf{w}_1$ and $\mathbf{w}_2$ at which convergence occurs are the optimum weights.

7- Since optimum $\mathbf{w}_1$ and $\mathbf{w}_2$ are obtained, the outputs $y_1(k)$ and $y_2(k)$ can be obtained.

8- Both $s_1(k)$ and $s_2(k)$ are obtained from $y_1(k)$ and $y_2(k)$.

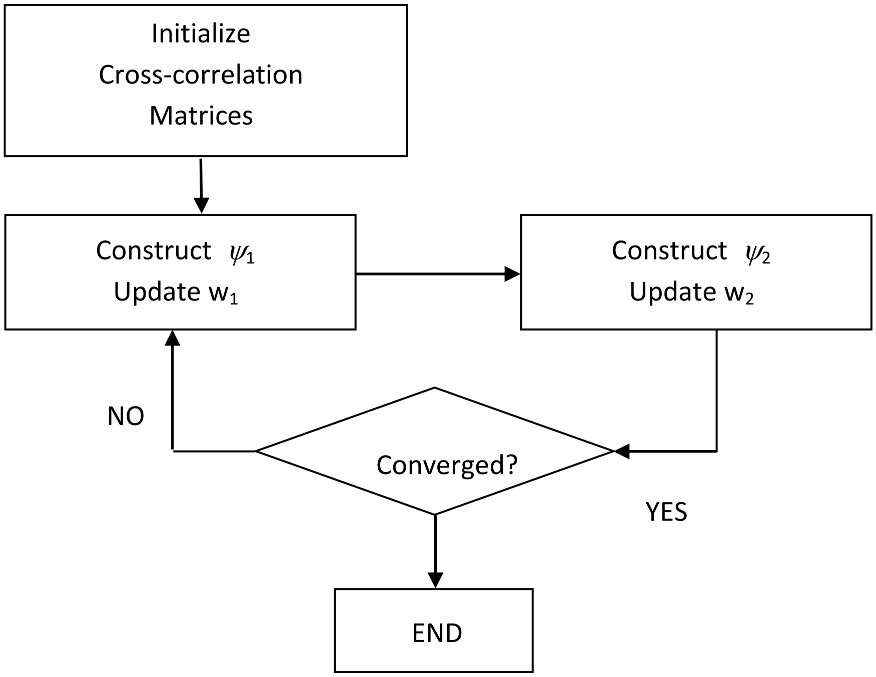


Fig. 3.4 Flowchart representing of the 2 × 2 algorithms.

CHAPTER 4

PROPOSED SPEAKER RECOGNITION SYSTEM BASED ON DEEP LEARNING

Most of the studies in speaker recognition consider an ideal scenario for training and testing without noise or reverberation. In real cases, some sources of noise may exist. Hence, in this thesis, we consider realistic cases comprising noise, reverberation, or interference in addition to the original signals. Hence, some pre-processing methods must be applied to reduce the degradations, and consequently improve the ASR performance. The ASR system behaves well for noise-free utterances. One of the most important challenges in the ASR systems is the problem of channel mismatch as the enrollment audio is gathered using one microphone and the test audio is produced by a different one.

It is important to note that the sources of mismatch vary, and are generally quite complicated. They are not limited to mismatches in the handset or recording apparatus, but they may include other combinations. There are some sources of recognition errors that may occur:

- mis-spoken or mis-read provoked expressions.
- severe emotional states (e.g. stress or compulsion).
- speaking style difference between training and testing.
- time-varying nature (intra- or inter-session) of microphone locations.
- poor or unreliable room acoustics (e.g. reverberation and noise)
- channel mismatch (e.g. using different microphones for enrollment and verification).
- different articulation of speech through the verification related to the training data.
- illness (e.g. head colds can alter the vocal tract).
- aging (the vocal tract can deviate from models with age).

Shortly, the main challenges of ASR systems are addressed here and their influence on the ASR implementation is studied.

4.1 Degradations Types

Three sorts of degradation are considered here: noise, reverberation, and interference.

4.1.1 Noise

Noise can be defined as any annoying signal that contaminates the speech signal causing hearing trouble and reduction in the intelligibility of speech signal. Noise appears in different formulas based on the transmitted information. The noise affects the ASR system performance, and hence affects the recognition rate. Noise is a random, undesirable signal that does not convey any useful information. It is any objectionable distortion in the speech signal that causes a reduction in the signal strength (SNR). Also, it results in poor intelligibility and poor hearing of speech. The sources of noise come from crosstalk or interference from other speakers [12].

The noise has a severe effect on the performance of speaker identification systems. Noise is supposed to be an additive signal that distorts the signal shape as shown in Fig. 4.1. The Additive White Gaussian Noise (AWGN) is a basic noise model used to mimic several random processes. It is a stationary random noise that has a uniform power spectrum across the frequency band of interest (same power for all frequencies). The noise samples have a Gaussian distribution in the time domain with zero-mean value, and the power spectral density is equal to the variance. While the additive colored noise is defined as an uncorrelated random noise process that does not have constant power spectral density. It has a Gaussian amplitude distribution. The noise power is not equally distributed over the whole spectrum, but localized in a specific range of the bandwidth [12].

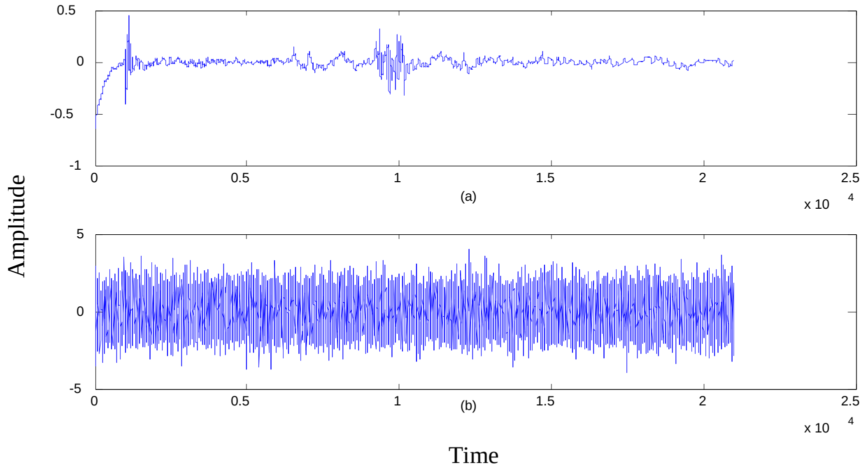


Fig. 4.1 Speech signal (a) clean speech (b) noisy speech.

4.1.2 Reverberation

The reverberation is a complex acoustical phenomenon. The reverberation effect is, in fact, some sort of multiple reflections with decaying energy as shown in Fig. 4.2. It degrades the intelligibility and quality of the speech signals, its effect on the spectrogram of speech signal is shown in Fig 4.3. A very important parameter that characterizes the reverberation is the reverberation time [54]. It is defined as the time taken by the signal to decay to 60 dB from its initial value at detection as shown in Fig. 4.4. The longer the reverberation time is, the more severity of the reverberation effect and the poor quality of the recorded speech signal. This degree of severity affects the further signal processing tasks applied to the speech signal.

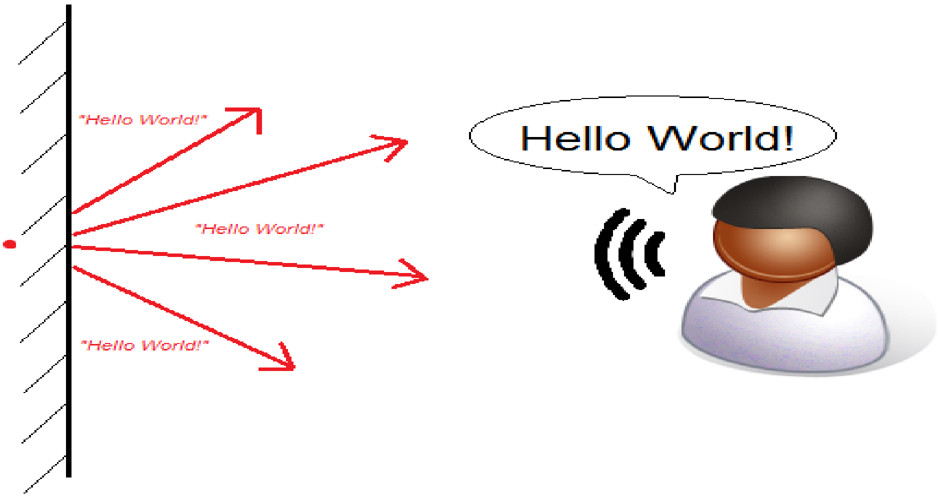


Fig. 4.2 Reverberation phenomena [54].

The reverberation leads to a severe effect on the performance of the ASR system. Normally, reverberant speech signals are recorded in closed rooms. These signals can be modeled as follows [55]:

$$z(t) = s(t) * h(t) \tag{4.1}$$

$$h(t) = a.\exp\left(-\frac{6.9t}{T_R}\right).n(t) \tag{4.2}$$

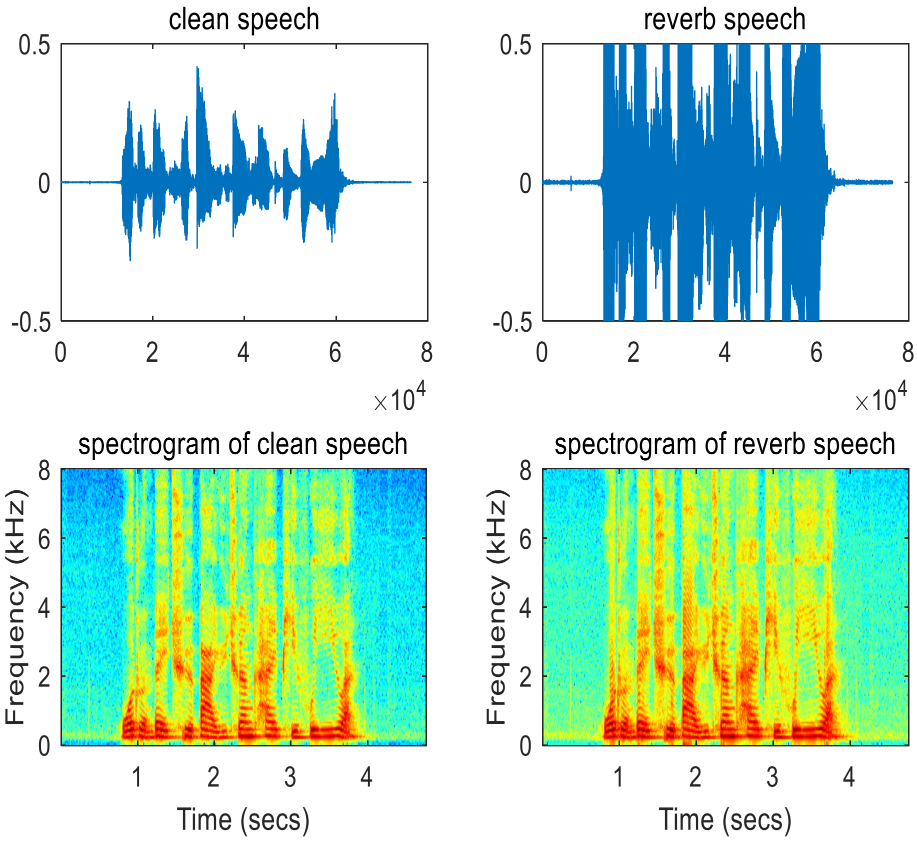


Fig. 4.3 The effect of reverberation on speech signals.

where $s(t)$ is the original speech signal, $h(t)$ is the room impulse response, $z(t)$ is the reverberant speech signal, a is a constant, t is the time, $n(t)$ is white noise and T_R is the reverberation time. The original speech signal is destroyed with long room impulse responses [56]. Reverberation of speech can also be modeled by a comb filter, which adds delayed versions of the signal to itself [57, 58].

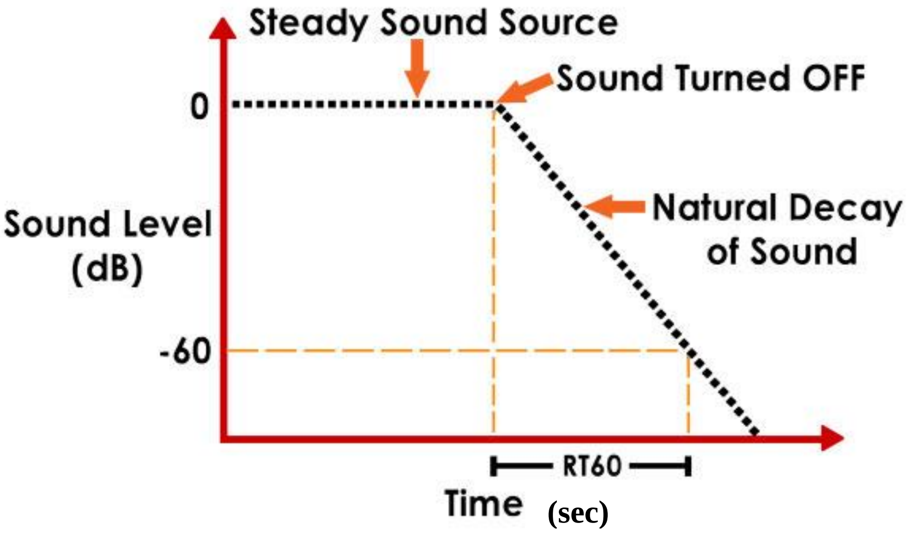


Fig. 4.4 Illustration of the reverberation time [54].

4.1.3 Interference of Speech Signals

Most studies in speaker recognition applications consider an ideal scenario for training and testing without interference. In real cases, some sources of interference may exist. Hence, signal separation needs to be applied to reduce the interference effect. Different techniques of blind signal separation are applied to improve the recognition rate of the text-independent speaker identification system. Blind signal separation is an important branch of signal processing as it deals mainly with mixed signals, which are frequently encountered in real life. Real-life signals are frequently mixed with undesired signals. This fact has motivated the evolution of blind signal separation algorithms [53].

Simulation results have revealed that, the ASR in the presence of interference leads to low identification scores. Signal separation has succeeded in enhancing these scores. So, the utilization of a signal separation algorithm is expected to enhance the performance of the ASR system. Blind signal separation for efficient text-independent SR systems is presented in this chapter. The word blind means that there is no a priori information about the mixed signals, and their sources. The mathematical model of the signal separation method was presented in Chapter 3 [50-53].

4.2 Speech Pre-processing

The main aspiration of speech enhancement is to improve the clarity and intelligibility of the speech signal. The key-point of speech enhancement is noise reduction, this disturbing noise is produced due to the background noise. There are some common enhancement algorithms such as spectral subtraction, wavelet denoising,… etc. Also, signal separation algorithm is used to overcome interference exists in speech signal. The mathematical model for the spectral subtraction, wavelet denoising, and signal separation are presented in details in Chapter 3.

4.3 Proposed Speech Classification Approach

An efficient approach for the classification of speech signals as reverberant or not is introduced in this section. The reverberation is a severe effect encountered in a closed room. So, it may affect subsequent processes or applications. It also deteriorates the speech processing system performance. So, it is important to classify the speech signal into either a reverberant or a clean speech signal, the process of classification is shown in Fig. 4.5. The spectrogram is utilized as an image generated from speech signals to be classified with a CNN. The spectrum and MFCCs are used as features to be classified by the LSTM-RNN. Two models are proposed and compared with a simple Benchmark model. Simulation results up to 100% classification accuracy are obtained for 100 utterance classification process, each of 36000 samples. This can help in performing an initial step in any speech processing system that comprises quality level classification. The performance of the proposed approaches were compared with that of the Benchmark model.

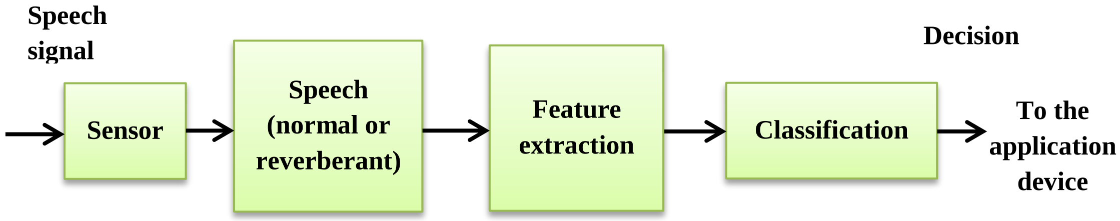


Fig. 4.5 The speech classification process.

The clustering technique employs two modes: training and testing modes. Each mode has two stages; extracting some distinguishing attributes and classification. In this thesis, the speech signal is used in its 2-D form. This can be implemented by obtaining the speech spectrogram. It is a good image representation of the speech signal [40]. Two models are used here for classification as shown in Fig 4.6: LSTM-RNN and deep CNN. The benefit of using a deep CNN is that it can make both feature extraction and classification. It uses some kernels that depend on the convolution process to extract some features called feature maps.

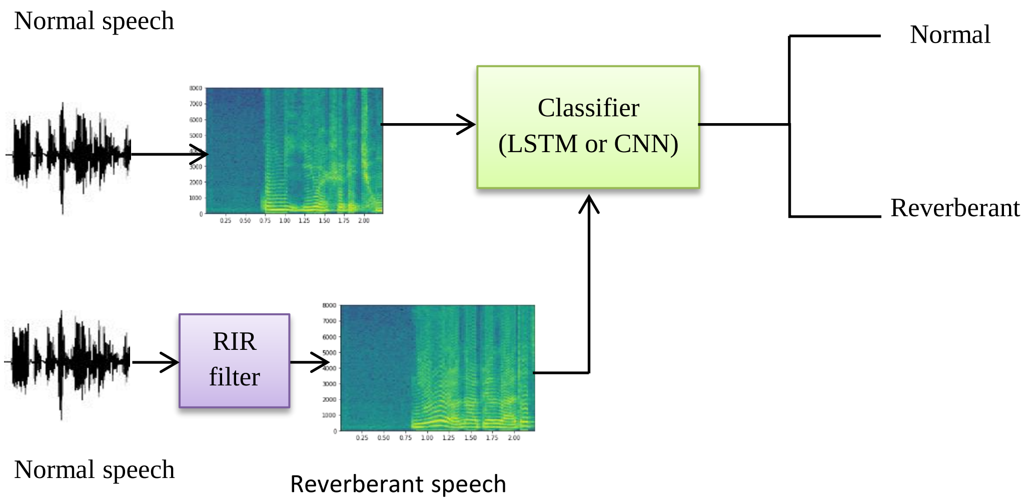


Fig. 4.6 Proposed speech classification approach.

The suggested approach is based on transforming the 1-D speech signal into its 2-D form, by obtaining its spectrogram as an image representation. In this proposed approach, there are two clusters of speech: normal and reverberant speech. Reverberant speech is modeled by using a Reverberant Impulse Response (RIR) filter with a reverberation time (RT60) equal to 0.5 sec. As mentioned above, clustering techniques have two stages: training and testing. In the training stage, a model based on the extracted features is made for each group and deposited in a database, while in testing, also a model is prepared and matched with that kept in the database to find the best match. In the LSTM-RNN, spectrum of speech is calculated and fed to the LSTM-RNN network as a 2-D feature matrix. In the CNN, the spectrogram is fed as an image and the network extracts the features by itself and the classification is performed by obtaining the final decision through the fully-connected layer.

4.3.1 Dataset Description and Results

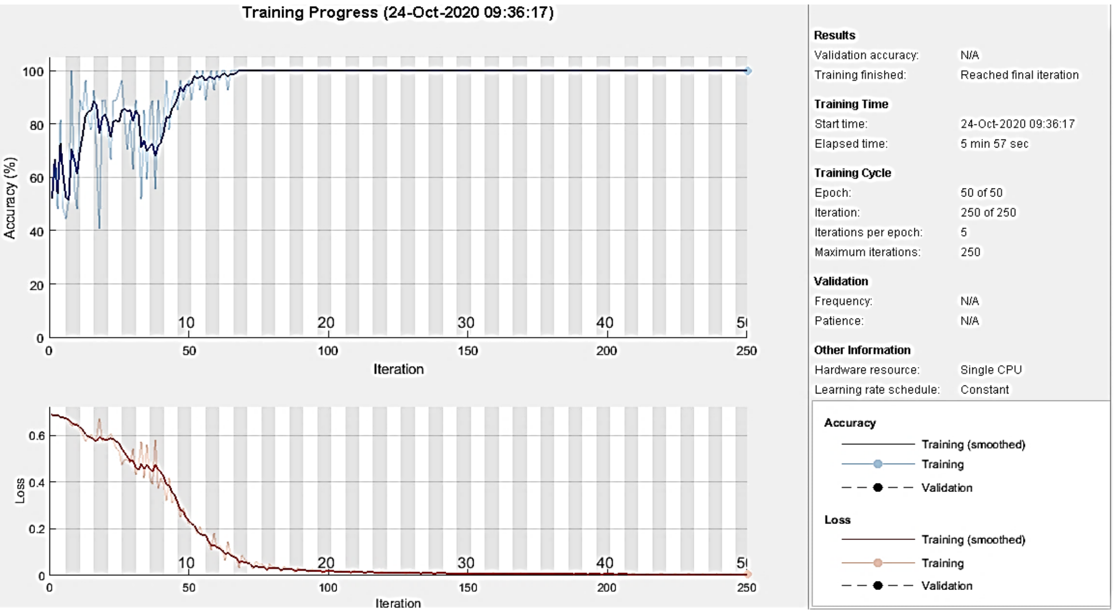
The used data in this dissertation is a subset of a much bigger dataset called the Chinese Mandarin Corpus dataset [59]. This dataset was recorded in silence in an in-door environment using cell phones. Ten speakers were considered in the simulations [five female and five male] with 100 utterances for each. The speech has a sampling frequency of 16,000 samples per second. The utterances are sub-divided into two parts: 70% utterances for training and the remaining 30% utterances for testing. The utterances are absolutely different in both training and testing stages. The speech signals are converted to a set of attributes known as features. These features are used as the input to the LSTM-RNN. A window of 25 milliseconds is considered, the 25 milliseconds window allows us to identify a precise time at which the signals change [18, 19]. The LSTM-RNN input layer size equals the number of input coefficients.

The features taken here are MFCCs, spectra and log-spectra. From every frame of size 256 samples, 13 coefficients, 129 coefficients and 129 coefficients are extracted in the cases of MFCCs, spectra, and log-spectra features, respectively. The whole feature vector enters the network at the same time, and each coefficient corresponds to a node in the input layer. The network works as a sequence classifier, not a frame classifier. The feature vectors from one speaker are seen as a sequence mapped to one target. A description of the used dataset is summarized in Table 4.1. In addition, the training progress of the LSTM-RNN and the CNN are shown in Fig. 4.7 and Fig. 4.8, respectively. The accuracy measures the system performance and reported in Table 4.2 and Fig 4.9, which is defined by Eq. 4.3:

$$Accuracy\ \% = \frac{Number\ of\ correct\ classification\ trials}{Total\ number\ of\ classification\ trials} \times 100\% \qquad (4.3)$$

Table 4.1 Dataset description [59]

Number of training utterances	100 [50 normal and 50 reverberant]
Number of testing utterances	30
Length and sampling rate of speech	36000 samples, and 16 kHz, respectively
Size of spectrogram image	224×224
Image type and format	RGB, and png
Software	Matlab R2017b
Kernel filter size	3×3
Number of filters for layers 1, 2, 3 and 4	16, 32, 64 and 128, respectively
Pooling size	2×2
RT60	0.5 sec



Results
Validation accuracy: N/A
Training finished: Reached final iteration

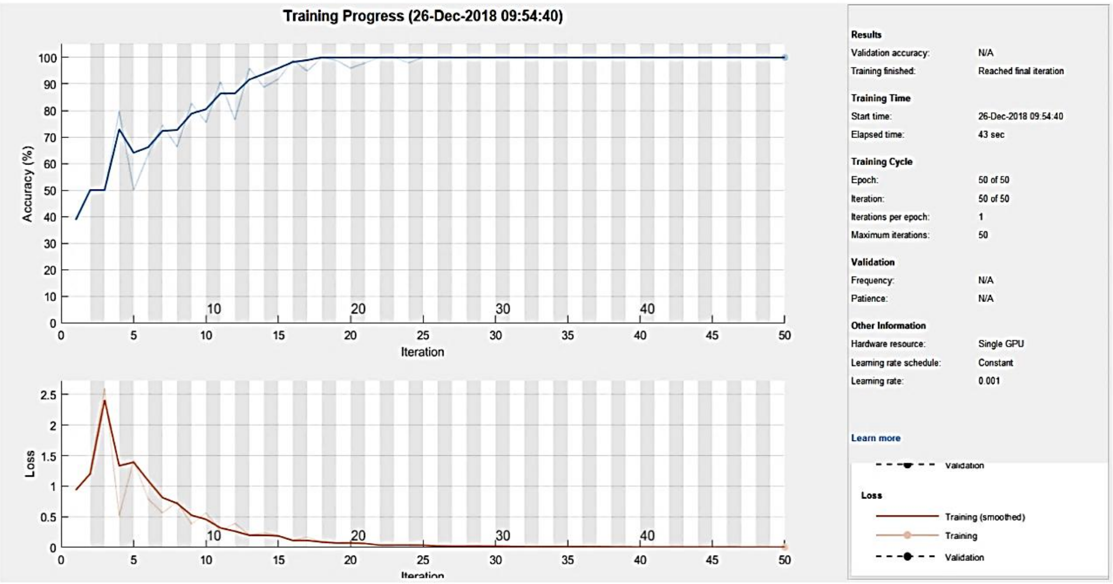
Training Time
Start time: 24-Oct-2020 09:36:17
Elapsed time: 5 min 57 sec

Training Cycle
Epoch: 50 of 50
Iteration: 250 of 250
Iterations per epoch: 5
Maximum iterations: 250

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant

Fig. 4.7 LSTM training progress.



Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 26-Dec-2018 09:54:40

Elapsed time: 43 sec

Training Cycle

Epoch: 50 of 50

Iteration: 50 of 50

Iterations per epoch: 1

Maximum iterations: 50

Validation

Frequency: N/A

Patience: N/A

Other Information

Hardware resource: Single GPU

Learning rate schedule: Constant

Learning rate: 0.001

Fig. 4.8 CNN training progress.

Table 4.2 Accuracy of classification using different methods

Method	Description	Accuracy %
LSTM-MFCCs	MFCCs	100
LSTM-Spectrum	Spectrum	100
Benchmark model	1-layer	92.86
CNN-2L	2-layer	96.43
CNN-3L	3-layer	100

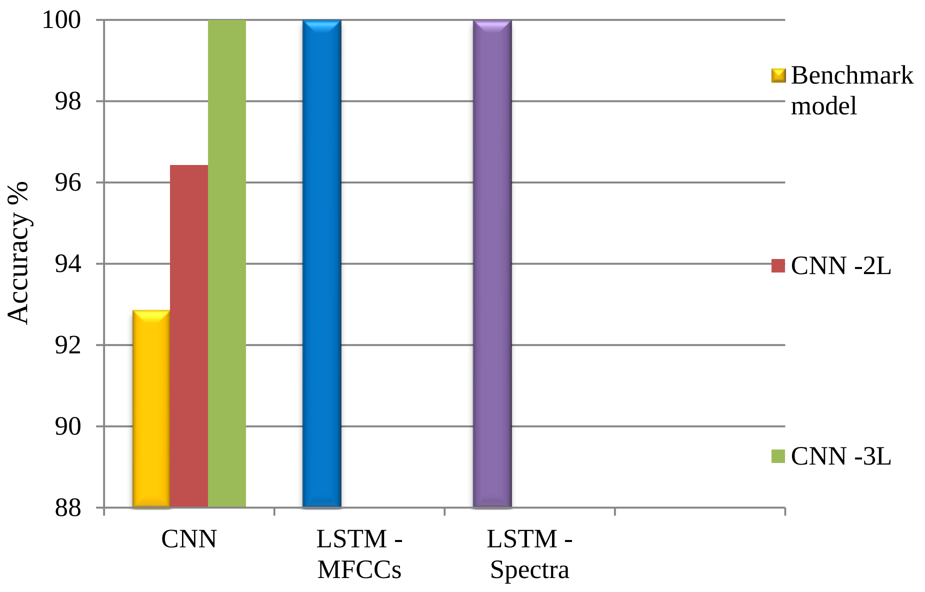


Fig. 4.9 Accuracy versus number of layers for the CNN and the LSTM-RNN.

Table 4.2 and Fig. 4.9 present the accuracy of the suggested approach in the case of using LSTM-RNN with both MFCCs and spectrograms, which reaches 100%. These results demonstrate the ability of the suggested approach to well distinguish between reverberant and normal speech. Also, it presents the accuracy of the Benchmark model and the CNN with two and three layers, which reaches 100% for the model of three layers. The performance of the two deep neural networks has been compared with that of the simple Benchmark model. Simulation results prove that a 100% accuracy for the quality level classification approach can be achieved with a 3-layer CNN. In addition, the MFCCs and spectra are used as features with LSTM to achieve an accuracy of 100%.

4.4 Proposed Text-independent Speaker Recognition Algorithm Using LSTM-RNN and Speech Enhancement

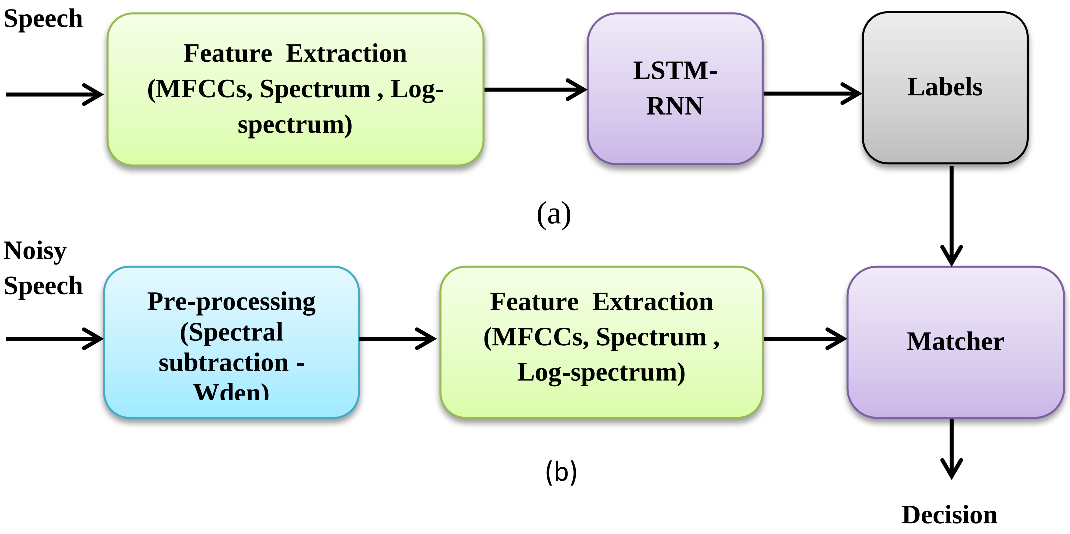
Speaker recognition revolution has led to the inclusion of speaker recognition modules in several commercial products. Most published algorithms for speaker recognition focus on text-dependent speaker recognition. In contrast, text-independent speaker recognition is more advantageous as the client can talk freely to the system. In this thesis, text-independent speaker recognition is considered in the presence of some degradation effects such as noise, reverberation, and interference. Mel-Frequency Cepstral Coefficients (MFCCs), spectra and log-spectra are used for feature extraction from the speech signals. These features are processed with the LSTM-RNN as a classification tool to complete the speaker recognition task. The network learns to recognize the speakers efficiently in a text-independent manner, when the recording circumstances are the same. The recognition rate reaches 95.33% using MFCCs, while it is increased to 98.7% when using spectra or log-spectra. However, the system has some challenges to recognize speakers from different recording environments. Hence, different speech enhancement techniques, such as spectral subtraction and wavelet denoising, are used to improve the recognition rate to some extent. The proposed approach shows superiority, when compared to the algorithm of R. Togneri and D. Pullella (2011) [19].

Text-independent speaker recognition is the much more stimulating than text-dependent speaker recognition. Speaker recognition systems have two stages: training and testing. In the training stage, a model for each speaker is created from a suitable representation of the speech created from the extracted features to discriminate between speakers [12]. Feature extraction is the most important step in speaker recognition systems. The MFCCs, spectra and log-spectra are used here for speaker representation. In the testing stage, also called classification stage, a model similar to that created in the training is made, and it is matched with all stored models.

Discrimination-based learning procedures are used in the classification process. In this thesis, LSTM-RNN, which is a specific RNN architecture, is used. In acoustic modeling, the LSTM-RNN is more effective than the Deep Neural Network (DNN). In contrary to feed-forward systems, RNNs are cyclic, and they can be translated as they have input through time steps. This makes them successful in learning of consecutive data, which gives them a type of crosswise memory over time. So, recurrent networks have two sources of input:

the present and the current past, which consolidate the decision about how they can react to new information.

The speech signal may be degraded due to some environmental conditions such as reverberation and noise. Consequently, some speech enhancement strategies are used to improve the speaker recognition system performance. The spectral subtraction [43] and the wavelet denoising [46] are used in a pre-processing stage before extracting the features in order to obtain robust features for good discrimination. Therefore, in this chapter, a proposed text-independent speaker recognition system is presented in the presence of some degradation effects such as noise and reverberation based on MFCCs and spectrum of speech as shown in Fig. 4.10. The features are handled with LSTM-RNN for classification. The achieved accuracy reaches 98.7% with spectrum using LSTM. The performance of the network is deteriorated, when the test utterances are degraded with noise and reverberation. This deterioration is the motivation for using some enhancement techniques to improve the recognition rate, which reaches 90% with the spectral subtraction method. The results of the proposed approach are compared to those obtained by R. Togneri and D. Pullella [19], as they use the MFCCs with Gaussian Mixture Models (GMMs).


Speech
Feature Extraction (MFCCs, Spectrum , Log-spectrum)
LSTM-RNN
Labels
Noisy Speech
Pre-processing (Spectral subtraction - Wden)
Feature Extraction (MFCCs, Spectrum , Log-spectrum)
Matcher
Decision

(a)

(b)

Fig. 4.10 Proposed speaker recognition system [(a) training (b) testing].

The proposed speaker recognition system is shown in Fig.4.10. It is based on recognizing speakers from any spoken phrases. The speech signals are processed and converted into sets of discrete features that can be used as input to the LSTM-RNN. The features taken here are MFCCs, spectra and log-spectra. From every frame of size 256 samples, 13 coefficients, 129 coefficients and 129 coefficients are extracted in the cases of MFCCs, spectra and log-spectra features, respectively. The LSTM-RNN is a powerful classifier that has been recently applied in the speaker recognition. One reason for the popularity of the LSTM-RNN is its good performance in classifying a sequence of data.

The distortion of speech signals can cause a poor recognition performance. In the testing phase, the degraded speech is first pre-processed using an enhancement method, and features are extracted and fed to the LSTM-RNN.

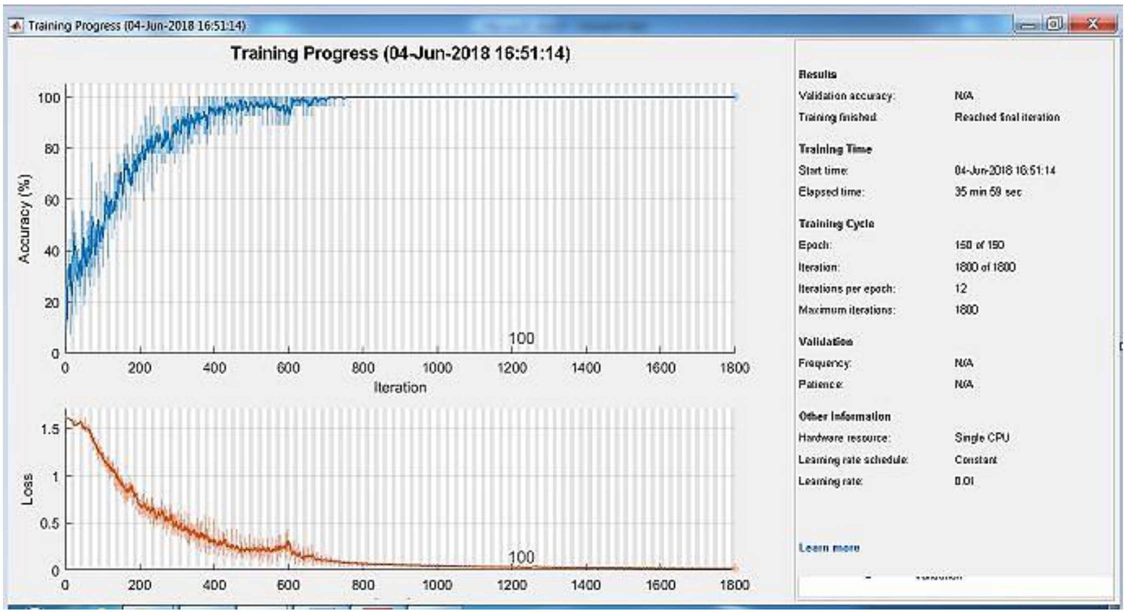
4.4.1 Simulation Results with LSTM-RNN and Speech Enhancement

Five female speakers with 500 utterances were considered in the simulations. The speech has a sampling frequency of 16,000 samples per second. The utterances are sub-divided into two parts: 350 utterances for training and the remaining 150 utterances for testing. The utterances are absolutely different in both training and testing. The sound waves were converted to sets of attributes known as features to be used as the input to the LSTM-RNN model. A window of 25 milliseconds was considered [18]. The LSTM-RNN input layer size equals the number of input coefficients. The network works as a sequence classifier not a frame classifier.

The quality of the input signal has a significant effect on the speaker recognition performance. Speech can be distorted due microphone motion, microphone change between training and testing or amplitude clipping, when the recording gain is too high. Noisy utterances are modeled by adding Additive White Gaussian Noise (AWGN) with different levels, and also reverberation is modeled by filtering with a certain impulse response as mentioned in previous sections. The reverberation time taken here is equal to 0.5 sec. The recognition rate is used to estimate the ASR system performance defined by Eq. 4.4.

$$Recognition\ rate\ \% = \frac{Number\ of\ success\ identification\ trials}{Total\ number\ of\ identification\ trials} \times 100\% \tag{4.4}$$

The effect of noise on the recognition performance was reported at different SNRs in table 4.4. Enhancement methods such as spectral subtraction and wavelet denoising have been used. The performance of the system with the undistorted utterances is satisfactory, but it gets degraded with the distorted utterances. It is enhanced, when the SNR is increased and enhancement methods are applied. The best accuracy is obtained when using spectral features with spectral subtraction at 30 dB. Generally, the LSTM-RNN is able to identify speakers well in text-independent mode and in the same recording scenario. The performance of the LSTM-RNN is degraded, when the test utterances are degraded with noise, and is improved by using some enhancement techniques. The training progress of the LSTM-RNN with the three types of features are shown in Figs 4.11, 4.12, and 4.13. Figures 4.14 and 4.15 show the change of the recognition rate with SNR for the three feature extraction methods with enhancement techniques. Tables 4.3, 4.4, 4.5, 4.6, and 4.7 report the recognition accuracy from undistorted utterances, distorted utterances, distorted utterances with wavelet denoising, distorted utterances with spectral subtraction and reverberant utterances, respectively. The results obtained are compared to those obtained by R. Togneri and D. Pullella [19]. The obtained results show the superiority of the spectrum features compared to MFCCs and log-spectrum features.



Results
Validation accuracy: N/A
Training finished: Reached final iteration

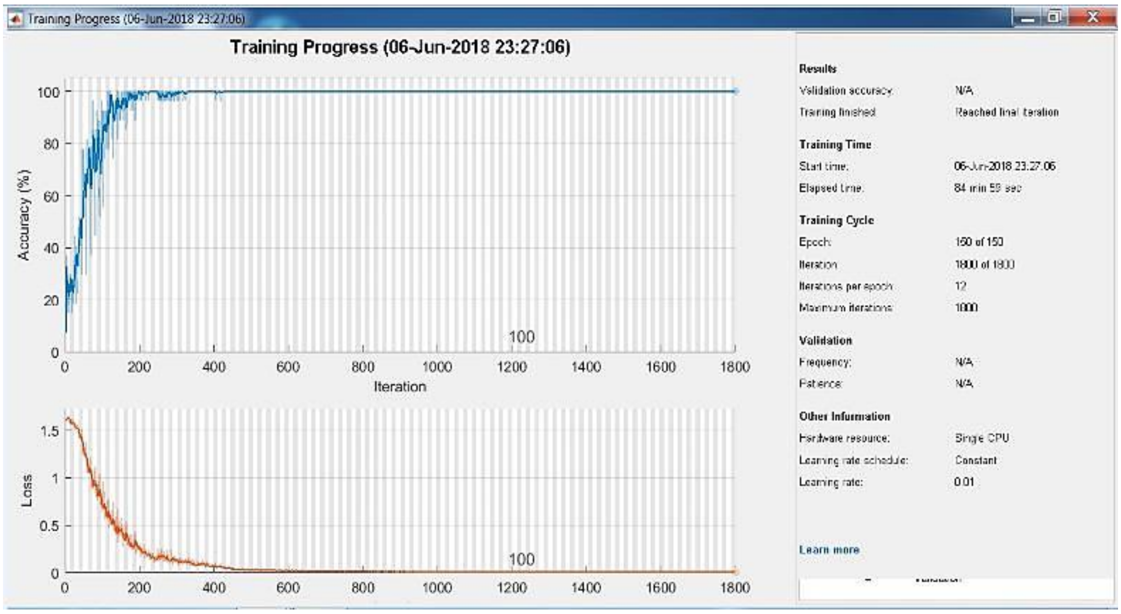
Training Time
Start time: 04-Jun-2018 16:51:14
Elapsed time: 35 min 59 sec

Training Cycle
Epoch: 150 of 150
Iteration: 1800 of 1800
Iterations per epoch: 12
Maximum iterations: 1800

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Fig. 4.11 Training progress of LSTM network (MFCCs).



Results
Validation accuracy: N/A
Training finished: Reached final iteration

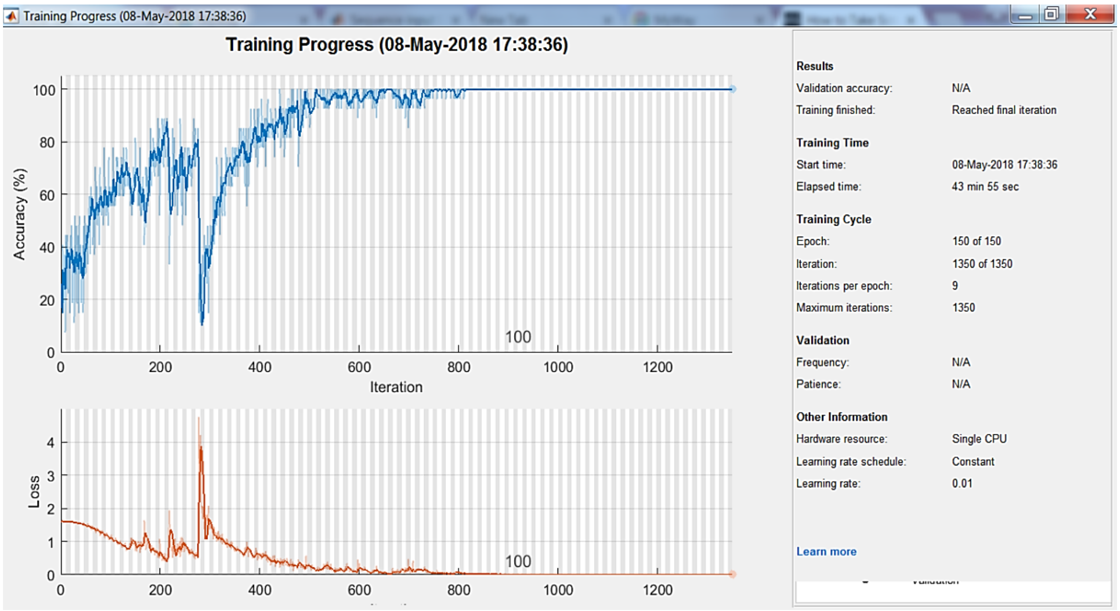
Training Time
Start time: 06-Jun-2018 23:27:06
Elapsed time: 84 min 59 sec

Training Cycle
Epoch: 150 of 150
Iteration: 1800 of 1800
Iterations per epoch: 12
Maximum iterations: 1800

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Fig. 4.12 Training progress of LSTM network (log-spectrum)



Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 08-May-2018 17:38:36

Elapsed time: 43 min 55 sec

Training Cycle

Epoch: 150 of 150

Iteration: 1350 of 1350

Iterations per epoch: 9

Maximum iterations: 1350

Validation

Frequency: N/A

Patience: N/A

Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.01

Fig. 4.13 Training progress of LSTM network (spectrum)

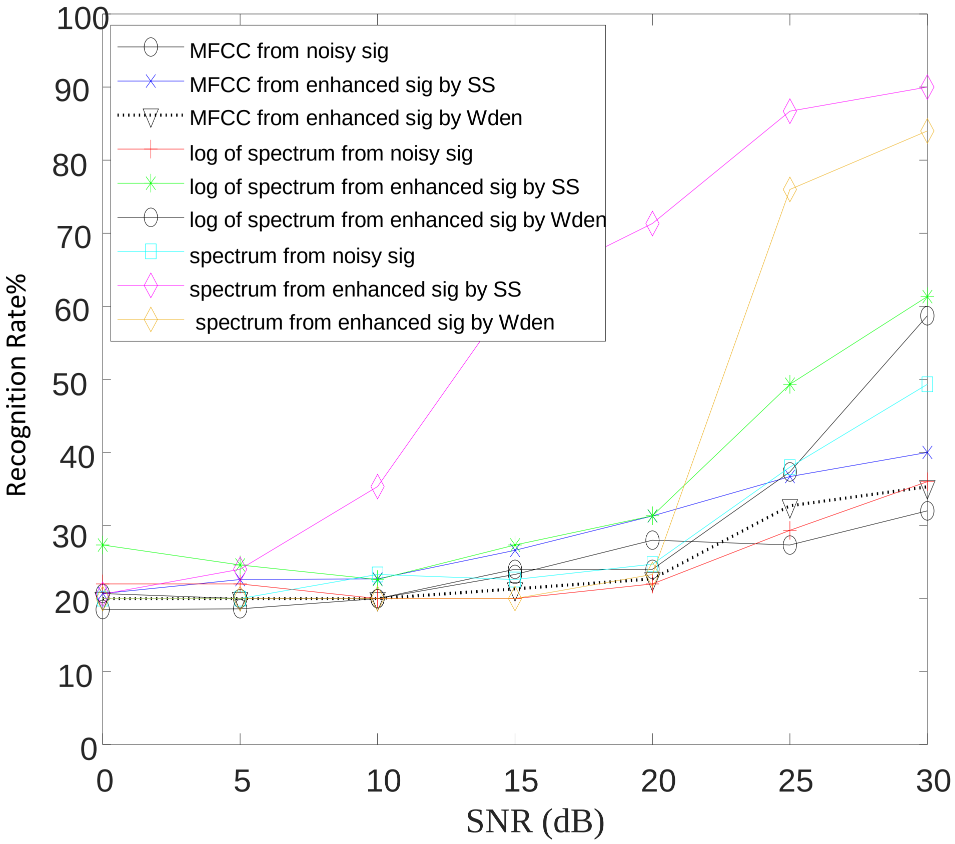


Fig. 4.14 Recognition rate versus SNR in dB.

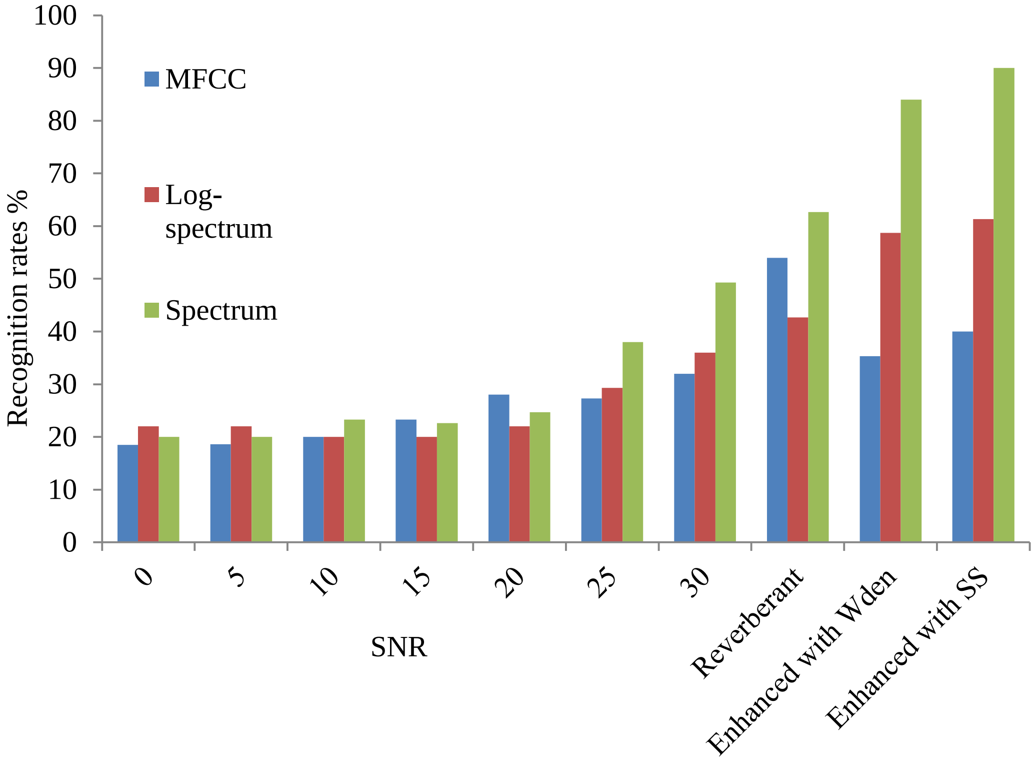


Fig. 4.15 Illustration of recognition rate versus SNR in dB.

Table 4.3 Recognition rate with undistorted utterances.

Feature Extraction Method	**Recognition rate %**
MFCCs	**95.33**
Log spectrum	**98.7**
Spectrum	**98.7**
MFCCs with GMM-UBM [19]	**94.5**

Table 4.4 Recognition rate with noisy utterances at different SNRs.

Feature Extraction Method	Recognition rate %						
	SNR=0 dB	SNR=5 dB	SNR=10 dB	SNR=15 dB	SNR=20 dB	SNR=25 dB	SNR=30 dB
MFCCs	18.5	18.6	20	23.3	28	27.3	32
Log-spectrum	22	22	20	20	22	29.3	36
Spectrum	20	20	23.3	22.6	24.7	38	49.3
MFCCs with GMM-UBM [19]	-	3.1	10.9	-	42.2	-	74.2

Table 4.5 Recognition rate after pre-processing with wavelet denoising

Feature Extraction Method	Recognition rate %						
	SNR=0 dB	SNR=5 dB	SNR=10 dB	SNR=15 dB	SNR=20 dB	SNR=25 dB	SNR=30 dB
MFCCs	20	20	20	21.33	22.7	32.7	35.3
Log-spectrum	20.7	20	19.3	24	24	37.3	58.7
Spectrum	20	20	20	20	23.3	76	84
MFCC with GMM-UBM [19]	-	3.1	10.9	-	42.2	-	74.2

Table 4.6 Recognition rate after pre-processing with spectral subtraction

Feature Extraction Method	Recognition rate %						
	SNR=0 dB	SNR=5 dB	SNR=10 dB	SNR=15 dB	SNR=20 dB	SNR=25 dB	SNR=30 dB
MFCCs	20.7	22.6	22.6	26.6	31.3	36.7	40
Log-spectrum	27.3	24.6	22.7	27.3	31.3	49.3	61.3
Spectrum	20.6	24	35.3	62	71.3	86.7	90
MFCCs with GMM-UBM [19]	-	3.1	10.9	-	42.2	-	74.2

Table 4.7 Recognition rate with reverberant utterances

Feature Extraction Method	Recognition rate %
MFCCs	54
Log spectrum	42.67
Spectrum	62.67

Text-independent speaker recognition under different conditions has been studied. The tools considered in this study are MFCCs, spectra and log-spectra features with an LSTM-RNN classifier. Also, speech enhancement was adopted to get the best performance, when utterances are degraded. The results ensure that the proposed LSTM-RNN is able to identify speakers well, in a text-independent mode, with undistorted utterances and in the same recording scenario with an accuracy that reaches 98.7%. For distorted utterances, a noticeable enhancement in the performance can be achieved, especially when using spectrum features with the spectral subtraction enhancement method leading to a 90% accuracy. These results were compared to those obtained by R. Togneri and D. Pullella to show the superiority of the proposed approach.

For distorted utterances at 20 dB for example, and using spectral subtraction with spectrum features, an accuracy of 71.3% was achieved, while R. Togneri and D. Pullella achieved an accuracy of 42.2%. Moreover, the effect of reverberation has been investigated and the recognition accuracy reaches 62.7% with spectrum features.

4.5 Performance Enhancement of Speaker Recognition with Noise Using Radon Transform

Speaker recognition under mismatching conditions is a challenging problem since robust feature extraction techniques are required. The LSTM-RNN is an efficient network that learns to recognize speakers well, text-independently, when the same recording circumstances are considered. When the operating conditions vary, the performance degrades. This section gives a new feature extraction method for ASR by RT and DCT. The spectrogram is efficient in demonstration and conveys information about audio features in the form of a pattern. In the suggested approach, the speaker's features are extracted by applying the RT on the pattern presented in the spectrogram. The RT has been used to get useful audio features from the speech spectrogram. Radon transform at a certain displacement adds up the pixel values in the image along a straight line in a specific direction [49].

The suggested approach computes Radon projections for thirteen orientations and captures the acoustic characteristics of the spectrogram [47]. The DCT is applied on the Radon projections, and it yields low-dimensional feature vectors. This technique is computationally efficient, text-independent, robust to variations and insensitive to additive noise. In this chapter, Radon projections of the spectrograms of the speech signals are used as features, since they are insensitive to noise and reverberation conditions.

4.5.1 Proposed Feature Extraction Method Based on Radon Transform

Radon transform is a good tool that picks up the directional features of an image. The spectrogram of the speech is obtained as an image representation of the power spectrum that is used with the RT. The Radon transform assembles the pixel intensity values in the spectrogram image along straight lines in certain directions with specific displacements.

The spectrogram assimilates acoustic features as energy, pitch, fundamental frequency, formants and time in the form of a pattern. The RT captures these features in the pattern by projecting it onto distinct orientation slices [47-49]. The Radon projection is obtained from the spectrogram of the speech signal, and then 2D DCT is applied. This technique improves the system recognition accuracy. It is text-independent and less sensitive to noise and reverberation. The proposed ASR system is shown in Fig. 4.16.

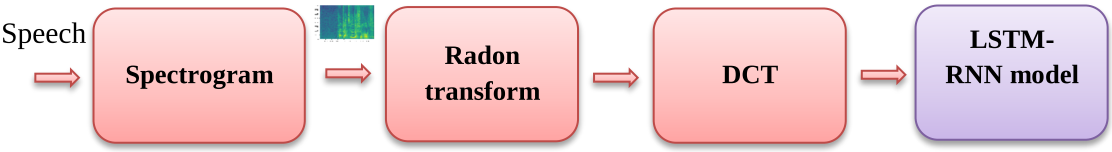


Fig. 4.16 Proposed ASR system based on Radon features.

Steps of the proposed approach:

- In the training phase, the spectrograms of speech signal are obtained as 2D images.
- The Radon projections of speech spectrograms are obtained.
- 2D DCT is performed on the Radon projections.
- The obtained features are used to train the LSTM-RNN.
- In the testing phase, the noisy utterance features are extracted by the same steps in used in the training and matched with the stored models.

The spectrogram is obtained for each wave using STFT with a window length of 256 samples. Thirteen orientations are used and at each orientation features are captured by Radon projection [47-49]. Only 50% of the features after applying DCT on the Radon projection are taken as it give the best performance. The input layer size of the LSTM-RNN net equals the number of input coefficients. The whole feature vector enters the network at the same time, each coefficient corresponding to a node in the input layer. The network works as a sequence classifier, not frame classifier since the feature vectors from one speaker are seen as a sequence mapped to one target.

The quality of the input signal has a significant effect on speaker identification performance. The effect of the noise on the performance of the ASR was reported at different SNRs. The recognition rate is used to measure the performance quality of the system. The AWGN is added to the speech corpus used in the testing phase at different SNRs from 0 to 30 dB. In addition, reverberation is added to testing utterances using RIR filter. Three feature extraction methods were used here: MFCCs, spectrum and the proposed feature extraction method based on RT. The recognition rates for the three cases are given in Table 4.8 and Fig. 4.17. Also, the training progress of the three cases are illustrated in Figs 4.18, 4.19, and 4.20.

Table 4.8 Recognition rate with the proposed features

Feature Extraction Method	**Recognition rate %**								
	Clean speech	**SNR= 0 dB**	**SNR= 5 dB**	**SNR= 10 dB**	**SNR= 15 dB**	**SNR= 20 dB**	**SNR= 25 dB**	**SNR= 30 dB**	**Reverberant speech**
MFCCs	**95.33**	**18.5**	**18.6**	**20**	**23.3**	**28**	**27.3**	**32**	**54**
Spectrum	**98.7**	**20**	**20**	**23.3**	**22.6**	**24.7**	**38**	**49.3**	**62.67**
Proposed Radon features	**84.67**	**25.33**	**36**	**45.33**	**58**	**72.67**	**80**	**82.67**	**80.67**

The recognition rate of the ASR from the proposed features based on Radon transform shows a greet enhancement from noisy utterance at different SNR. Since the Radon transform is insensitive to any variation of the speech signal, so it is good tool to represent the speech signal in noisy environment. Also, in case of reverberant utterance the proposed features has the superiority than the other two feature extraction methods. The essence that the proposed features outperforms the two other feature extraction methods in case of noisy and reverberant utterance.

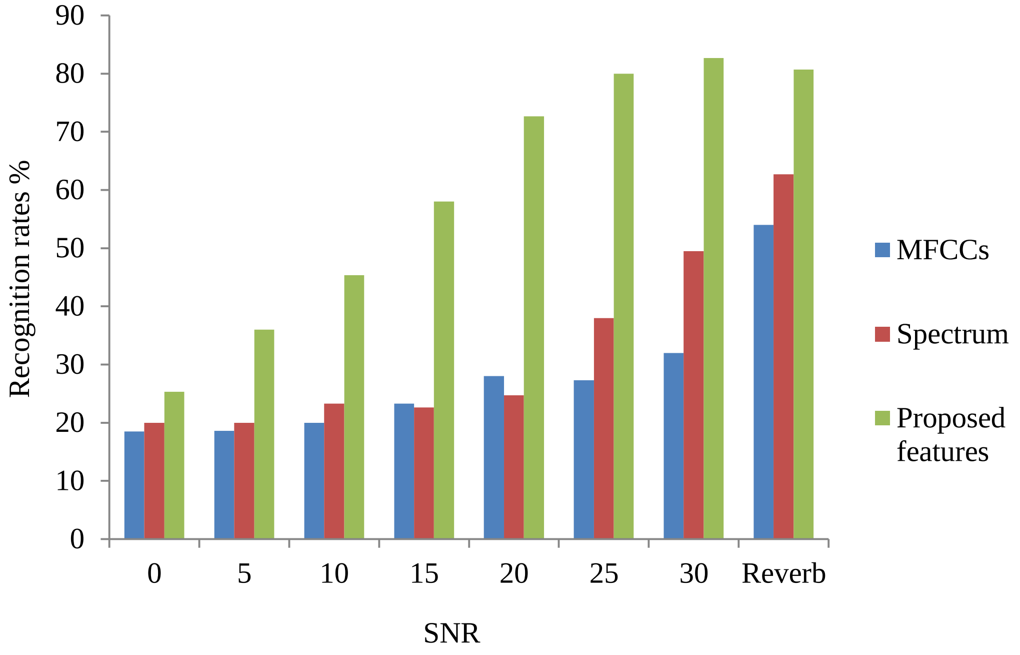
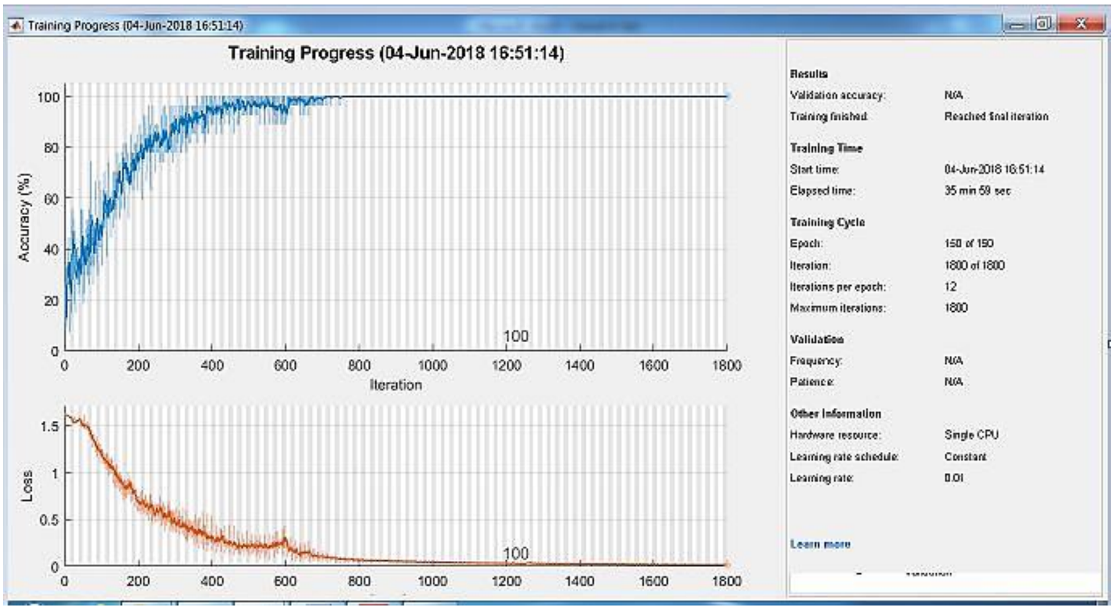


Fig 4.17 Recognition rates versus SNR and reverberation.



Results

Validation accuracy:	N/A
Training finished:	Reached final iteration

Training Time

Start time:	04-Jun-2018 16:51:14
Elapsed time:	35 min 59 sec

Training Cycle

Epoch:	150 of 150
Iteration:	1800 of 1800
Iterations per epoch:	12
Maximum iterations:	1800

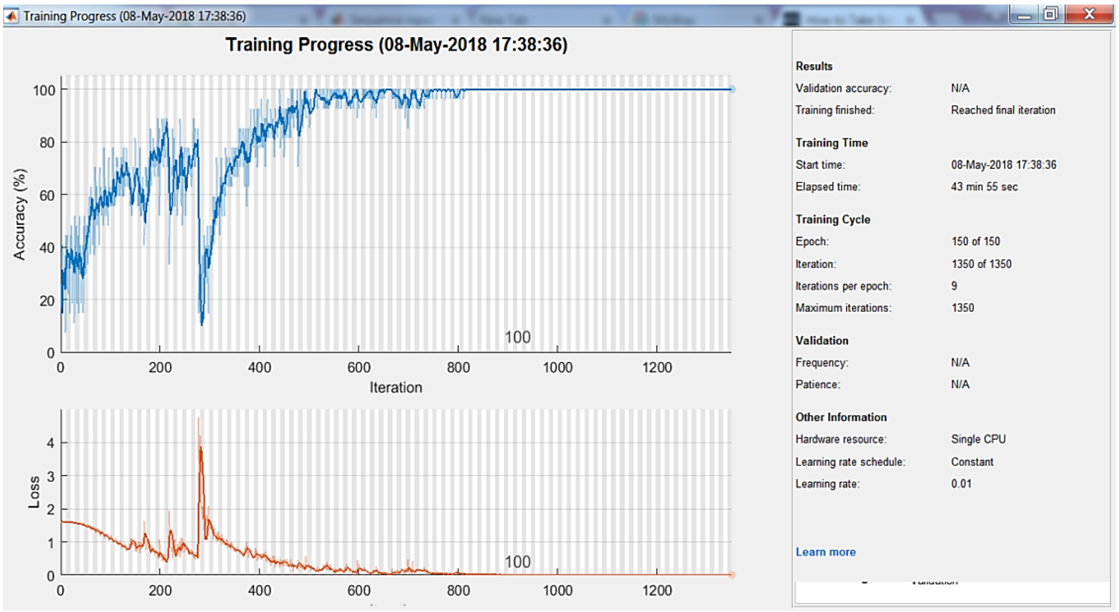
Validation

Frequency:	N/A
Patience:	N/A

Other Information

Hardware resource:	Single CPU
Learning rate schedule:	Constant
Learning rate:	0.01

Fig 4.18 Training progress in case of using MFCCs.


Training Progress (08-May-2018 17:38:36)

Training Progress (08-May-2018 17:38:36)

Results
Validation accuracy: N/A
Training finished: Reached final iteration

Training Time
Start time: 08-May-2018 17:38:36
Elapsed time: 43 min 55 sec

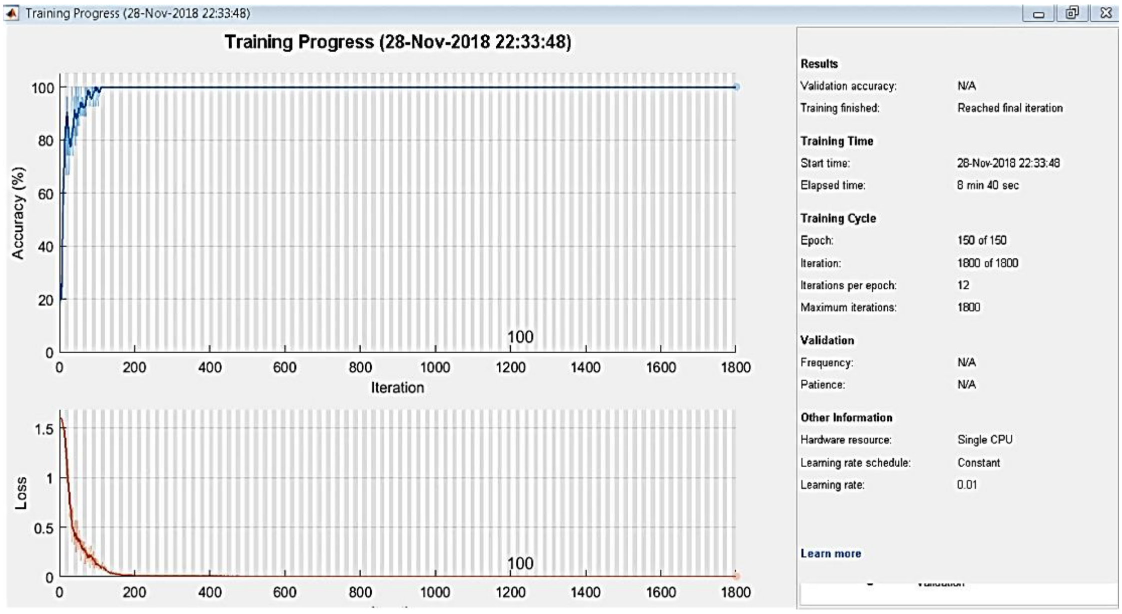
Training Cycle
Epoch: 150 of 150
Iteration: 1350 of 1350
Iterations per epoch: 9
Maximum iterations: 1350

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Learn more

Fig 4.19 Training progress in case of using spectrum.


Training Progress (28-Nov-2018 22:33:48)

Training Progress (28-Nov-2018 22:33:48)

Results
Validation accuracy: N/A
Training finished: Reached final iteration

Training Time
Start time: 28-Nov-2018 22:33:48
Elapsed time: 8 min 40 sec

Training Cycle
Epoch: 150 of 150
Iteration: 1800 of 1800
Iterations per epoch: 12
Maximum iterations: 1800

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Learn more

Fig. 4.20 Training progress in case of using proposed features.

Experimental results prove the superiority of the proposed approach in the case of noisy and reverberant utterances compared to the other two methods. The proposed approach is less sensitive to distortion than other approaches. Moreover, the training time is less than the training time taken by the other methods.

4.6 Effect of Reverberation on the ASR Based on CNN

This proposed approach is based on building an efficient deep learning model that can recognize speakers. This approach is based on the utilization of a CNN with different numbers of layers as shown in Fig. 4.21. Also, the effect of changing the pooling size and the filter size on the recognition process was studied. There are really two operation cases, the first case for utterances without reverberation and the second for the reverberation case. The used models consist of a convolutional (CNV) layers followed by a max pooling layers for the two cases. In both cases, the output from the convolution layer passes through the ReLU function, followed by the pooling layer. The pooling layer employs the mean pooling process of two-by-two sub-matrices. Finally, a global average pooling is used. Table 4.9 shows the model summary for each layer and its output shape.

Table 4.9 Model summary

Layer Type	Output Shape
CNV	(222, 222, 16)
Pooling	(111, 111, 16)
CNV	(109, 109, 32)
Pooling	(54, 54, 32)
CNV	(52, 52, 64)
Pooling	(26, 26, 64)
CNV	(24, 24, 128)
Pooling	(12, 12, 128)
Global Average Layer	(256)
Dense	(26)

The input is 224×224 pixel images. A CNN with layers having 16, 32, 64, 128, and 256 filters for layers 1, 2, 3, 4 and 5, respectively. Finally, a dense layer with a size of 10 is used for the classification task. A batch-size of 128 is used [40]. The learning rate is 0.001. The classification neural network consists of a single hidden layer and an output layer. Since we have 10 classes, the output layer is constructed with 10 nodes as shown in Fig. 4.22.

This model is carried out on 500 images for training and 150 images for testing from a 10 speaker dataset. A ten speakers were used (5 female and 5 male), each speaker has 65 utterances. The proposed model consists of CNV layers followed by a ReLU activation then a max-pooling layers. Finally, a global average pooling is used. The input to the network are images of 224×224 pixels, number filters of 16, 32, 64, 128, 256 for layers one, two, three, four, five respectively. Finally, a dense layer with a size of 10 is used for the classification task. The number of epochs are 50.

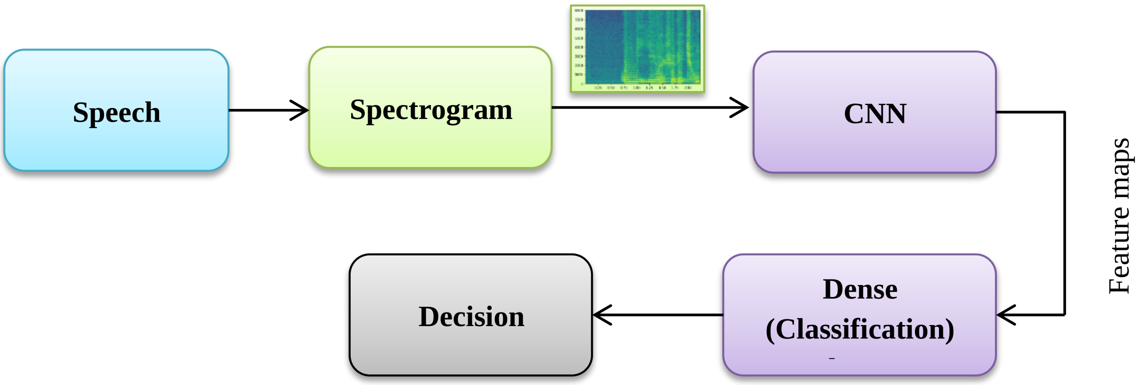


Fig. 4.21 Proposed ASR based on CNN.

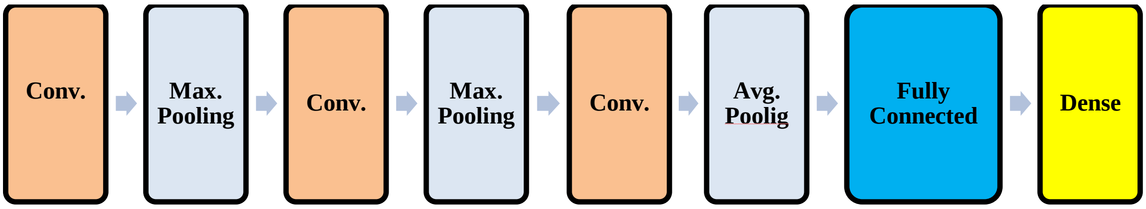


Fig. 4.22 CNN layers.

The proposed models results were compared with those obtained by Yanick Lukic et al. [40]. They use the spectrogram as an input to the CNN, and then study the design of the network for identification and clustering of 10 speakers. They use a kernel size of 4×4, max-pooling size of 4×4, stride size of 2×2 and a batch-size of 128. Their achieved accuracy reaches 97%. The recognition rate of the proposed CNN models shows the strength of the three and four-layer CNN model compared to the Yanick Lukic model as shown in Table 4.10.

Table 4.10 Recognition rates for different models without reverberation

Model description **Epoch = 50** **Pooling, kernel**	**Recognition rate %**					
	One layer CNN (1L)	**Two layer CNN (2L)**	**Three layer CNN (3L)**	**Four layer CNN (4L)**	**Five layer CNN (5L)**	**Yanick Lukic model [40] CNN (K=4×4, P=4×4)**
***P*=2×2 , *K*=3×3**	**95**	**95.6**	**96.25**	**96.88**	**98.12**	**97%**
***P*=2×2 , *K*=5×5**	**94.37**	**96.25**	**97.5**	**98.12**	**97.5**	
***P*=4×4 , *K*=3×3**	**93.13**	**95**	**97.5**	**98.75**	**96.8**	
***P*=4×4 , *K*=5×5**	**96.25**	**93.13**	**96.25**	**98,12**	**98.12**	

Table 4.11 Recognition rates for different models with reverberation

Model description **Epoch = 50** **Pooling , Kernel**	**Recognition rate %**				
	One layer CNN (1L)	**Two layer CNN (2L)**	**Three layer CNN (3L)**	**Four layer CNN (4L)**	**Five layer CNN (5L)**
***P*=2×2 , *K*=3×3**	**74.6**	**78**	**82**	**82.7**	**76.67**
***P*=2×2 , *K*=5×5**	**73.33**	**79.33**	**80.7**	**82.7**	**82.7**
***P*=4×4 , *K*=3×3**	**76**	**84**	**76.7**	**70.7**	**60.7**
***P*=4×4 , *K*=5×5**	**70.7**	**76**	**80**	**80**	**71.33**

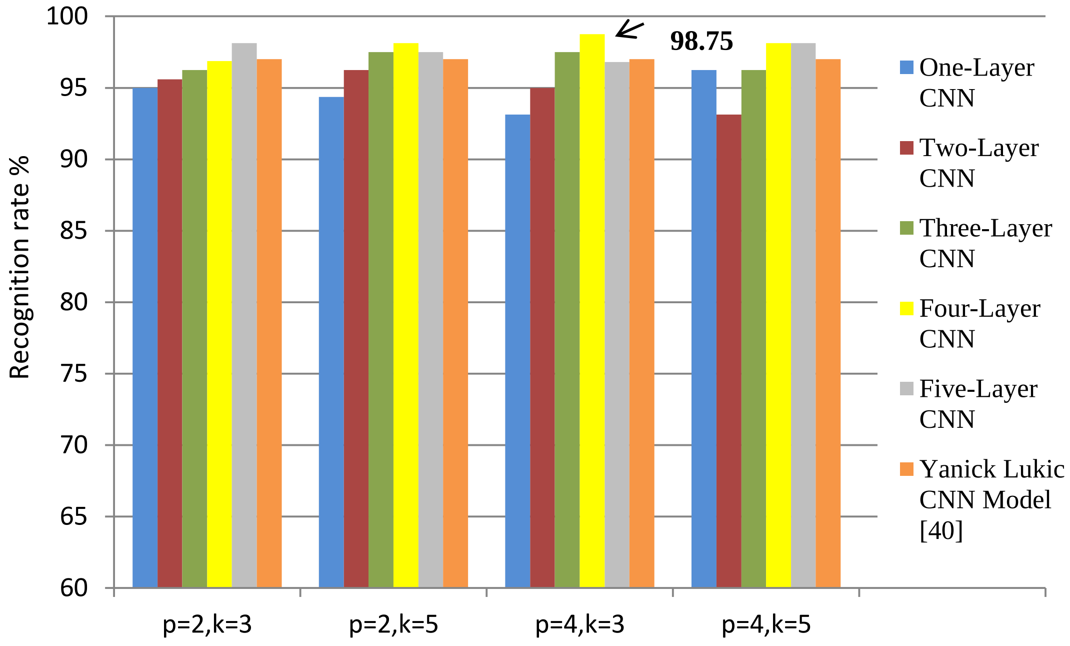


Fig. 4.23 Results summary without reverberation

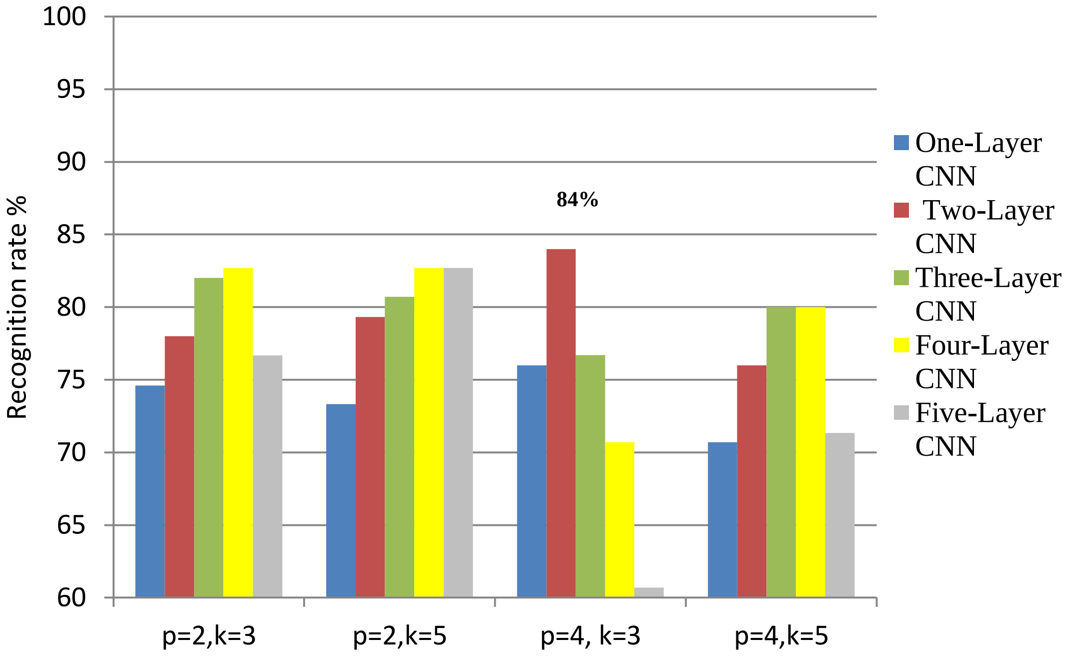


Fig. 4.24 Result summary with reverberation.

The proposed approach is concerned with the feasibility of using features generated from CNNs to recognize speakers using different models of CNN in text-independent recognition case. Also, the effect of reverberation phenomena has been investigated, which is noticeable on the recognition accuracy. This proposal adopts the speech utterances to benefit from deep learning with a CNN, by converting the speech signal into an image suitable for operation with a. Different CNN models for speaker identification have been presented with and without reverberation effect. The simulation results showed that the proposed models have superiority in performance compared to that of Yanick Lukic CNN model [40]. The results are obtained for different pooling sizes, kernel sizes and numbers of layers. The model of four layers, 4×4 pooling, and 3×3 kernel size has the best accuracy for clean speech, while the model of two layers, 4×4 pooling, and 3×3 kernel has the best accuracy in the case of reverberant utterances.

4.7 Proposed ASR System Based on Signal Separation and Deep Learning

Most researches in speaker recognition applications consider an ideal scenario for training and testing without noise or interference. In real cases, some sources of interference may exist. So, in this chapter, we consider a realistic case of interference with the original signals. Hence, signal separation needs to be applied to reduce the interference effect. Different techniques of blind signal separation are applied to improve the recognition rate of the text-independent speaker identification system. Blind signal separation is an important branch of signal processing, as it deals mainly with mixed signals, which are frequently encountered in real life. Real-life signals are frequently mixed with undesired signals. This fact has motivated the evolution of blind signal separation algorithms [53].

Simulation results have revealed that speaker recognition in the presence of interference leads to low identification scores. Signal separation has succeeded in enhancing these scores. So, the utilization of a signal separation algorithm is expected to enhance the performance of speaker recognition systems. Blind signal separation for an efficient text-independent speaker recognition system is presented in this chapter. The word blind means that there is no a priori information about the mixed signals and their sources [51-52].

The utilization of a signal separation algorithm is expected to boost the performance of a speaker recognition system. In order to achieve a satisfactory accuracy in many speaker recognition applications, a clean single-source input is required. So, different techniques of blind signal separation are applied to improve the recognition rate of the text-independent speaker recognition system. Mainly, this chapter is devoted to studying the text-independent speaker recognition process assisted with signal separation and deep learning techniques. The main objective of this chapter is to build a text-independent speaker recognition system and show how interference can degrade the ASR performance. Blind signal separation is used to improve the system performance. Separation is performed in time, DST and DCT domains. The common objective is to segregate and estimate the source signals without knowledge of the characteristics of the transmission channel. A 2×2 signal separation system is used. It is based on using the output decorrelation as the signal separation criterion. The mathematical model of the separation method is discussed in Chapter 3. Then, the segregated speech signals are used in the speaker recognition system, whose performance is compared to the conventional system with clean and mixed utterances without separation. [50-53].

The simulation results prove the effectiveness of the signal separation on the ASR system performance. Also, separation in the time domain has superiority compared to separation in the DST and DCT domains. Two models of deep learning are used; LSTM-RNN and CNN. Separation is performed in different domains and the results are compared to those of the conventional system with clean and mixed utterances without separation. Results reveal that separation, especially in the time domain, achieves higher recognition rates than those in other domains. The recognition rate reaches 97.33% with the three-layer CNN and 96% with the LSTM-RNN.

4.7.1 System Architecture

The first proposed speaker recognition system is shown in Fig. 4.25. The clean speech signals with known speakers are used in the training stage to train the LSTM-RNN classifier by the spectrum as features. The LSTM-RNN makes a label for each speaker, and this label kept as a reference. In the testing phase, the unknown speakers' signals mixed with undesired signals are first pre-processed by the blind separation method in order to obtain the source signals without interference. Then, features are extracted and supported to the LSTM-RNN classifier. The classifier matches the unknown utterances with each reference stored from training. In Fig. 4.25, the mixed speech signal is first pre-processed by the separation method, the spectrum is calculated in the form of a feature vector that is handled by the LSTM-RNN, which is a sequence classification technique.

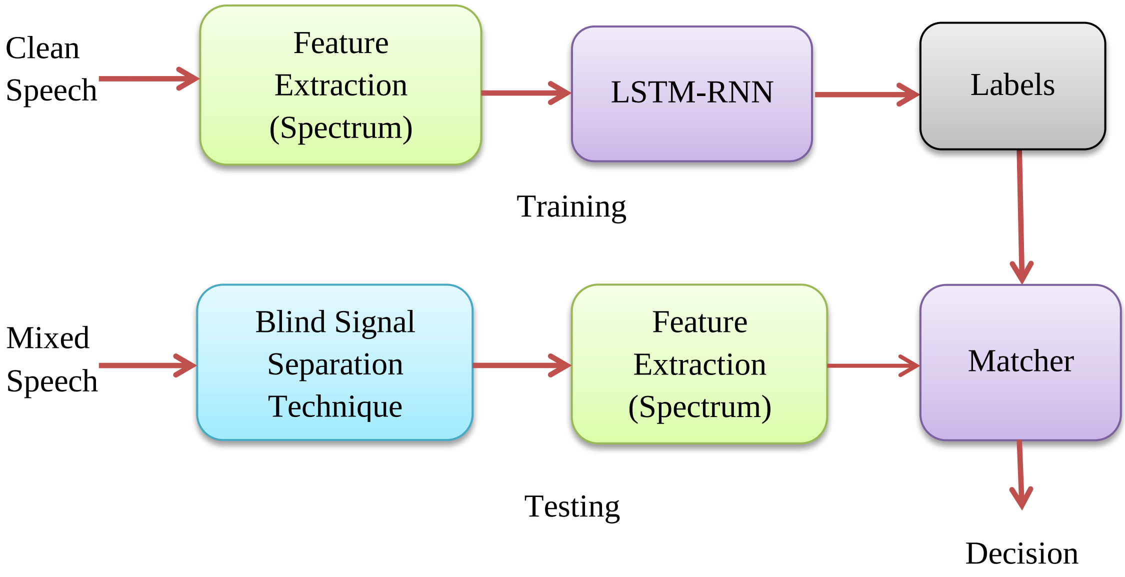


Fig. 4.25 Proposed ASR system based on LSTM-RNN and signal separation.

The second proposed SR system is shown in Fig. 4.26. The CNN is used here which has two networks: one for extracting features from the spectrogram images and the other is fully-connected layer for classification. Similar to the first proposed system, the CNN is first trained by clean speech signals for known speakers. The spectrograms are used as input images. The CNN extracts features from the spectrograms by means of kernel filters, and the features are handled by a fully-connected layer to map these features to labels. In the testing phase, the mixed speech signals are first pre-processed by the separation methods and the spectrograms are obtained as images to represent speech signals.

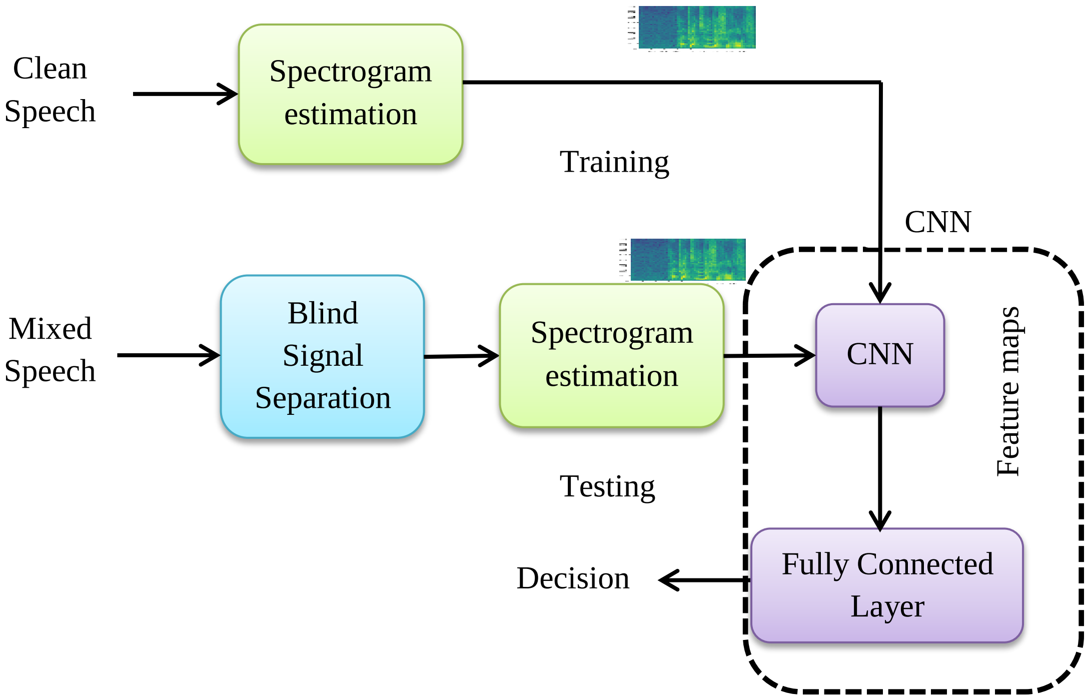


Fig. 4.26 Proposed ASR system based on CNN and signal separation.

For the two proposed models, the pattern matching system is first trained with clean utterances, and a label for each speaker is created and stored. In training phase, there are two states: clean utterances and mixed utterances. For clean utterances, the performance of the SR system is good, while for mixed utterances, the system performance degrades. Separation is used as a pre-processing tool on mixed utterances to ameliorate the system performance.

Separation is performed in time, DST and DCT domains, and the recognition performance is compared with the performance in cases of clean and mixed speech signals without separation. The steps of the proposed ASR system are as follows:

- Each speaker has a number of utterances. These utterances are sub-divided into training and testing segments (70% for training and 30% for testing).
- In the training phase, features are extracted and used to train the classifier.
- The classifier takes these features to make a model for each speaker. It associates each speaker's model to a label.
- In the testing phase, when an unknown speaker enters the system, a model is created by extracting features.
- The pattern matching method compares the unknown model with the known models and the matching score is obtained.
- Based on the matching score, a decision is made about the speaker's identity.

4.7.2 Simulation Result and Discussion

The same dataset is used here. Spectrogram is obtained for each wave using STFT with a window length of 256 samples. To obtain spectrum features of speech, the speech signal is first divided into frames of a fixed size and overlap of 25% or 50%. These frames have been arranged column-wise to form a matrix if the speech signal is a one-dimensional signal. Speech is first divided into frames of 256 samples each with an overlap of 50% between consecutive frames. The obtained 129 frames are then arranged column-wise to form a matrix of dimension 129×256. Discrete Fourier Transform (DFT) is then applied on this matrix column-wise. The spectrogram is plotted as the absolute magnitude of the transform matrix.

129 features are obtained by applying spectrogram from speech corpus and used with the LSTM-RNN of input layer size equal to the number of input coefficients. The whole feature vector enters the network at the same time, and

each coefficient corresponds to a node in the input layer. The network works as a sequence classifier, not frame classifier since the feature vectors from one speaker are seen as a sequence mapped to one target. Figure 4.27 reveals the effect of interference on the speech waveform. The speech is mixed with music. The figure also displays the effect of separation methods on the waveform of speech.

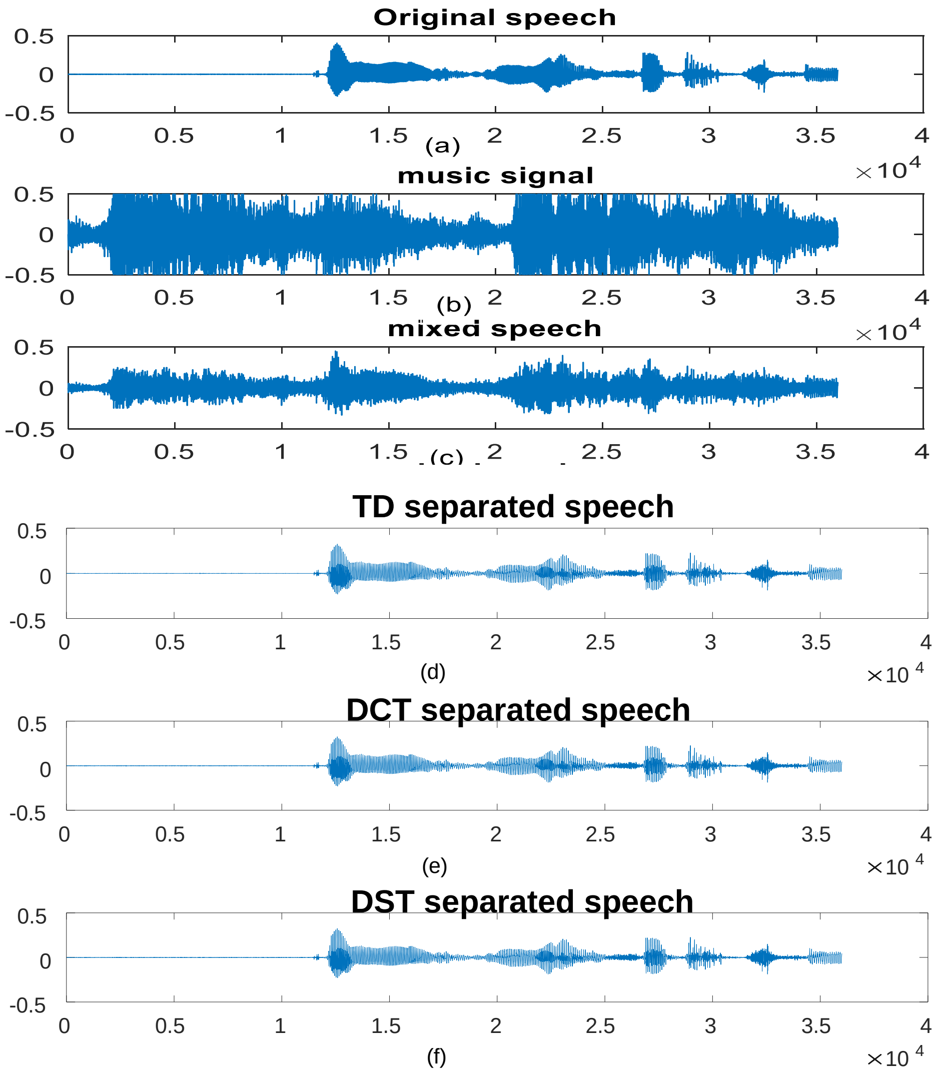


Fig. 4.27 Effect of interference and separation on speech signal (a) Original speech, (b) Music signal, (c) Mixed speech, (d) Speech separated in time domain, (e) Speech separated in DCT domain and (f) Speech separated in DST domain.

The figure displays separation in time, DST and DCT domain of the mixed speech waveforms. For CNN, speech signals are converted into spectrograms in the form of 2-D images of size 224×224. The number of kernel filters are 16, 32, 64 and 128 for layer 1, 2, 3 and 4, respectively, each of size 5×5. Also, a 2×2 max-pooling is used. Fig. 4.28 denotes the recognition rates for all methods, which are also noted in Table 4.12.

Table 4.12 Recognition rates for different separation methods

Model description Epoch = 50	**Recognition rate %**				
	Mixed speech	**Time separation**	**DCT separation**	**DST separation**	**Clean speech**
2L-CNN	**23.33**	**95.33**	**93.33**	**92.67**	**96.25**
3L-CNN	**34**	**97.33**	**93.33**	**94**	**97.5**
4L-CNN	**36.67**	**96.67**	**96**	**96**	**98.12**
5L-CNN	**37.33**	**94.7**	**94**	**93.3**	**97.5**
LSTM-RNN	**21.3**	**96**	**96**	**96**	**98**
Benchmark model	**29.33**	**94.67**	**91.3**	**91.3**	**94.33**

The effect of blind signal separation on the ASR system is shown in Fig. 4.28. Also, the training progress of the LSTM and CNN are shown in Fig. 4.29 and Fig 4.30. These figures reveal that signal separation of the desired speech signals is very important for robust speaker recognition, because interfering signals destroy the distinguishing features of speech signals.

A proposed text-independent speaker recognition from mixed speech using signal separation as a pre-processing step with deep learning is presented. The proposed approach is based on the blind separation of the speech signals in time and different domains including DST, and DCT. This obtained results ensure the feasibility of using features generated from the CNN for the ASR system. It adopts the speech utterances to benefit from deep learning with CNNs by converting the speech signals into images appropriate for the CNN to deal with. In addition, the LSTM was used as a classifier with spectrum used as features and the recognition rate reaches 96%, which is mostly better than the recognition rate in the case of mixtures that is 21.3%. The simulation results prove that the separation of the mixed speech signals has a noticeable impact on the text-independent speaker recognition system. For example, in the case of a CNN of three layers, the recognition rate from the mixed speech is 34%, while it reaches 97.33% after applying the separation, which is relatively close to the recognition rate in the case of clean utterances.

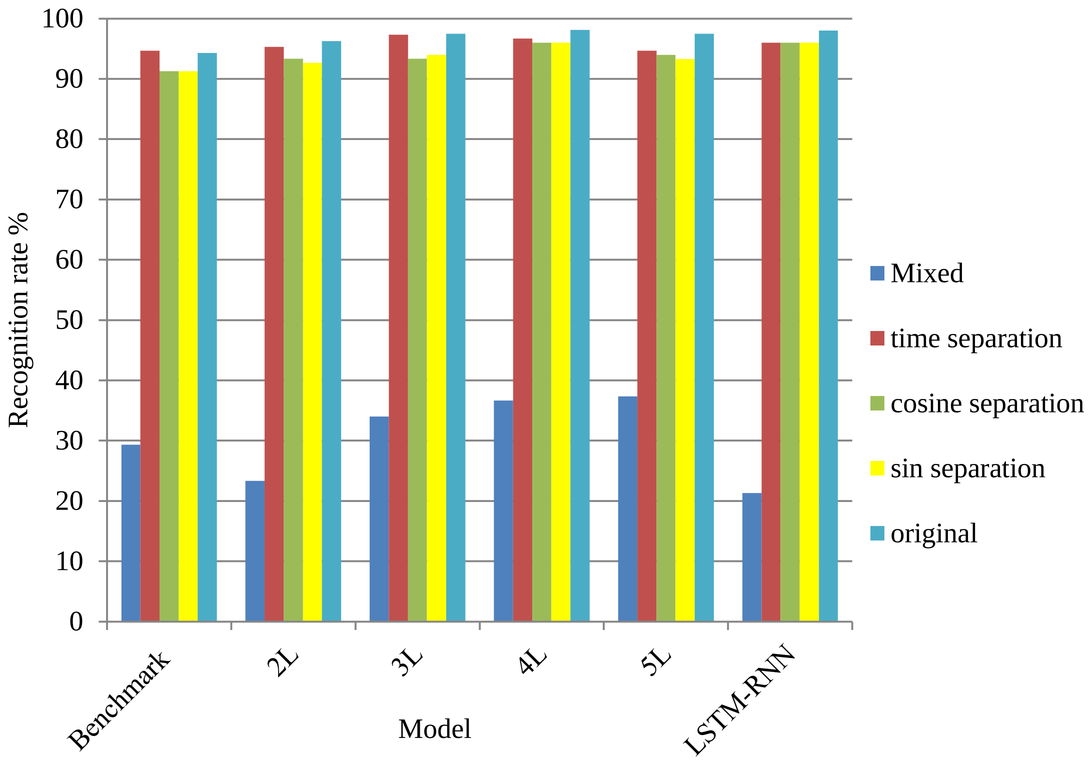
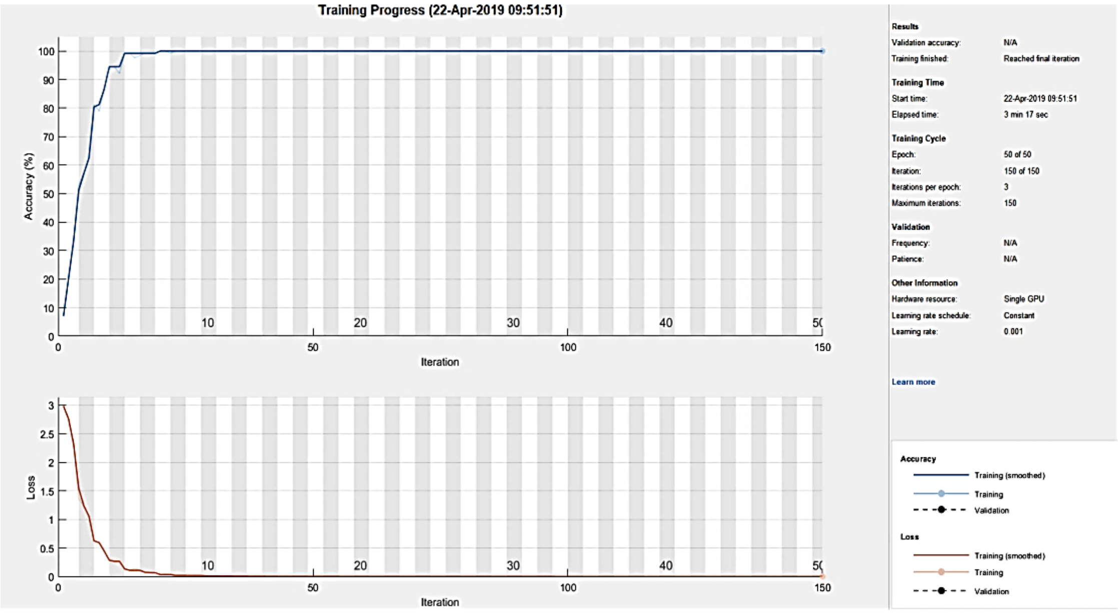


Fig. 4.28 Recognition rates for different models using different methods



Training Progress (22-Apr-2019 09:51:51)

Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 22-Apr-2019 09:51:51

Elapsed time: 3 min 17 sec

Training Cycle

Epoch: 50 of 50

Iteration: 150 of 150

Iterations per epoch: 3

Maximum iterations: 150

Validation

Frequency: N/A

Patience: N/A

Other Information

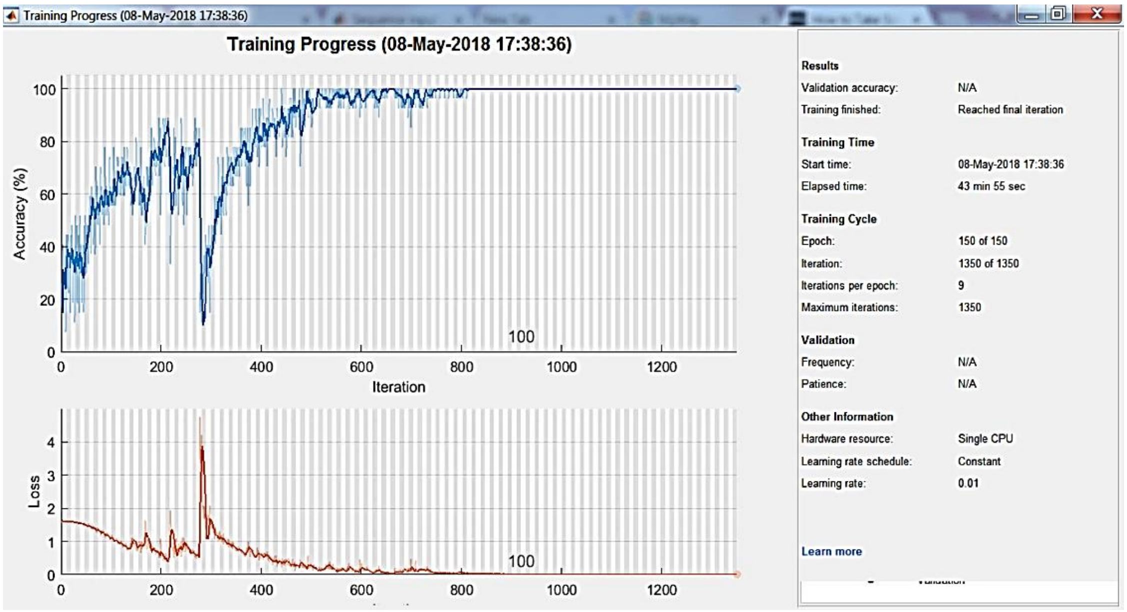
Hardware resource: Single GPU

Learning rate schedule: Constant

Learning rate: 0.001

Learn more

Fig. 4.29 CNN training progress



Training Progress (08-May-2018 17:38:36)

Training Progress (08-May-2018 17:38:36)

Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 08-May-2018 17:38:36

Elapsed time: 43 min 55 sec

Training Cycle

Epoch: 150 of 150

Iteration: 1350 of 1350

Iterations per epoch: 9

Maximum iterations: 1350

Validation

Frequency: N/A

Patience: N/A

Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.01

Learn more

Fig. 4.30 LSTM-RNN training progress.

A CNN with 2, 3, 4 and 5 layers is used. A comparison is held among the different models with and without the application of the blind signal separation algorithms mentioned before (time domain, DCT domain, and DST domain). A comparison study is held between the method with and without the application of the blind signal separation algorithms mentioned before. Simulation results reveal that the separation is an important processing stage prior to the ASR system, since the system performance degrades mainly in the presence of interference. The separation in the time domain is superior to separation in other domain, especially for the 3-layer CNN model as the recognition rate reaches 97.33%.

CHAPTER 5

CANCELLABLE SPEAKER RECOGNITION

5.1 Introduction

The biometric pattern is non-revocable and vulnerable to confidentiality attacks. Cancellable biometrics has been presented to resolve these issues. In this chapter, we present an approach to generate cancellable templates from speech signals that can be replaceable and can be re-released, when compromised. The proposed approach allows the generation of variable speech templates without likability. Speaker recognition is the process of identifying speakers by obtaining a voice signature for everyone regardless of the protection of the user's privacy. Since the voice is the used biometric, it is important to save the patterns from being attacked, and keep the user's confidentiality [60]. Cancellable speaker recognition is a new notion addressed for these problems.

In the proposed approach, the speech signals are converted into an images by obtaining their spectrograms to deal with a deep CNN. Only, a patch of the spectrogram is used as an input to the network based on a user-specific key. If the voiceprint is disclosed, this key can be changed to select another patch and prevent the original speech biometric from being compromised.

In this chapter, a CNN is used for both extracting features and pattern matching. Since the voiceprint is related to the person and cannot be substituted, it represents great risk if it is compromised. In order to solve the problem of template abuse and safeguard user’s secrecy, a new notion is tackled to resolve these problems called cancellable templates [61].

A proposed cancellable speaker recognition approach is introduced. This approach is based on converting the speech into an image, by finding its spectrogram. Then, only a patch of this spectrogram is used based on a certain key. This key is determined by the user. This means that one can create different signatures for different application. If the key is identified, the user can create a different key to guarantee safety and renewability of the speech biometric [60-63]. A deep convolutional neural network (CNN) is used here, since it is a good tool that can deal with images. The CNN can do both feature extraction and classification. So, there is no need for a pre-processing step for extracting features. The CNN has two networks: one for extracting features and another for classification. The simulation results reveal that the suggested approach is practical and it satisfies the desired criteria such as renewability, security, and performance. The accuracy of the proposed scheme reaches 98.75% with a CNN of three layers.

5.2 Cancellable Template and Points of Attacks

Commonly, a biometric system stores biometric patterns for recognition purposes. The stored biometric patterns in a database can be stolen by an impostor. An impostor can change the information in a template to imitate it as an actual user to make illegal actions. The main problem in traditional biometric system is the non-replaceable characteristic. This means that as soon as the template is stolen, it is considered unusable, and no new template can be re-released. Furthermore, the biometric templates are not recommended to be used in several applications. If a single biometric template is used in numerous applications, when one of these patterns is compromised, the user has to stop all associated applications.

Recently, the development of cancellable biometrics improved the biometric template protection. Cancellable biometrics generation depends on changing the biometric templates into other formats that can be changed when compromised (reusability property). The transformed biometric templates achieve the diversity property. The cancellable biometric system must achieve a performance comparable to that of a traditional speaker recognition system [62, 63].

The speaker recognition system is roughly drawn in Fig. 5.1. It consists of a device for collecting voice (sensor), a feature extractor unit, a matcher, a database, and an application device, which is controlled by the matcher output. The voice is collected by a sensor, and distinct features are obtained. A model based on the features is generated for each user in such a way that it is simply readable and comparable through the matching. During the authentication stage, the unknown input query is matched with the stored models. Matching is achieved by a learning-based classifier be able to classify the authorized and unauthorized persons. A matching score is obtained by finding the degree of similarity between the new and the stored biometric models. A decision is made based on the matching score [60-64].

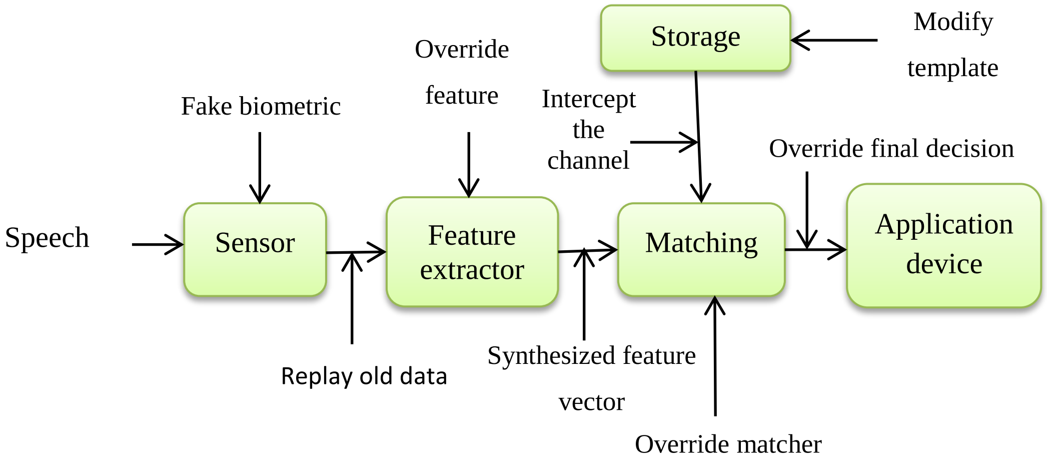


Fig. 5.1 Speaker recognition system and points of attacks.

There are many sources of attacks as illustrated in Fig. 5.1. The matcher can be attacked to revoke the last decision of the system. Protection schemes are advanced to create easy substituted or revoked biometric templates. These schemes are called cancellable biometric schemes. Cancellable biometric schemes must satisfy certain criteria such as renewability: the possibility to withdraw the pattern, security: the difficulty getting the original biometric data from the stored secure patterns and performance: the accuracy of the system should not be sensitive to template replacement. There are many cancellable biometric schemes that depend on random projection, bio-convolving, and non-invertible transforms.

5.3 Proposed Cancellable Speaker Recognition Scheme

In this chapter, a proposed protection scheme for speaker recognition based on spectrogram patch selection is used as shown in Fig. 5.2. First, the speech utterances from all users are transformed into spectrogram. Spectrograms are introduced to a CNN that perform both feature extraction and pattern matching. The spectrogram is a Time-Frequency Representation (TFR) of the signal, which embraces the complete characteristics of an audio signal in both spectral and temporal domains. The spectrogram is obtained by applying the Short Time Fourier Transform (STFT) on the signal.

By segmentation of the signal into frames of fixed length, a window with a little overlap is adopted. The spectrogram is the squared magnitude of the STFT as mentioned in Chapter 2.

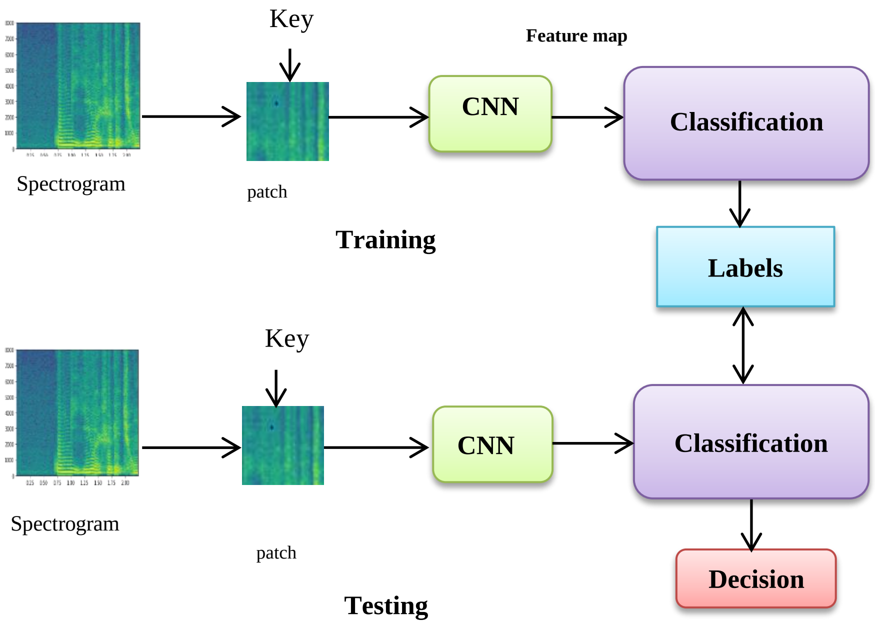


Fig. 5.2 Proposed cancellable speaker recognition system (training and testing)

The cancellable pattern is built up by only taking a patch of the spectrogram based on a certain key for each user, and then fed to the CNN as an input. As stated above, the CNN has two networks: one for extracting features (CNV) and the other for classification. A CNN with different layers was used to obtain the best performance. Comparisons are presented between different models. For a speech signal $s(n)$, the image spectrogram is obtained as $f(m,n)$, only a patch is taken based on a user-specific key. The transformed image is $f(i,j)$, where $i < m$, $j < n$. If the key is stolen or lost, it can be changed and a new one is extracted from the same spectrogram. Also, different keys can be used for different applications.

5.4 Simulation Results

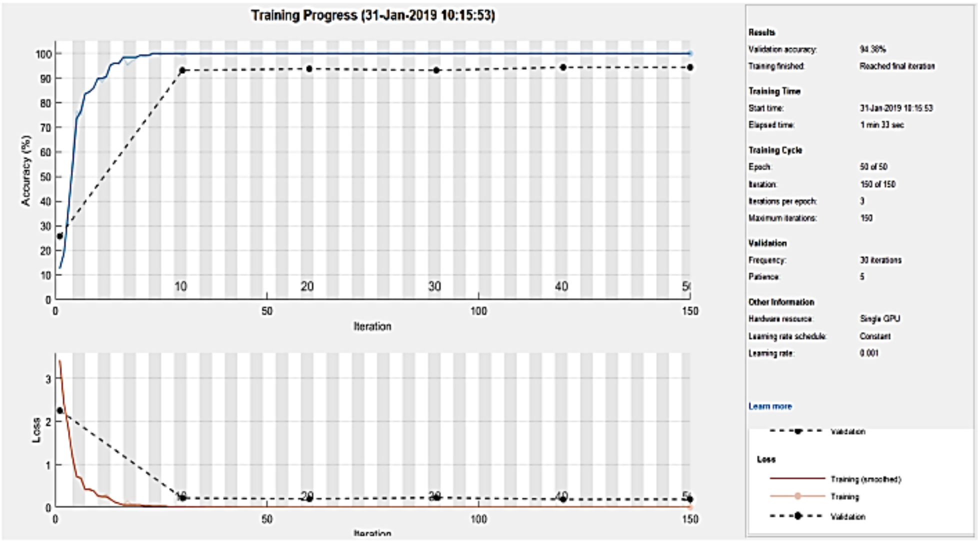
The same Chinese Mandarin Corpus dataset is used here. 10 speakers signals are used (5 female and 5 male). Each speaker has 65 utterances. The utterances are absolutely different in both training and testing. The speech has a sampling frequency of 16,000 samples per second. The speech signals are mapped to spectrogram images. There are 500 images for training and 150 images for testing. Also, the recognition rate is used to measure the performance of the system.

The simulation results are obtained for the four models of CNN; Benchmark model, two, three and four-layer models. The first model is the Benchmark model that consists of one CNV layer followed by one max-pooling layer. The second model consists of two CNV layers followed by two max-pooling layers. The third model composed of three CNV layers followed by three max-pooling layers. The fourth model has four CNV layers followed by four max-pooling. In every case, the output from the convolution layer passes through the ReLU function, followed by the pooling layer. The pooling layer employs the mean pooling process of two-by-two sub-matrices. Finally, global average pooling is used.

The input is a 161×161 pixel image. The CNN layers have a number of filters: 16, 32, 64 and 128 for layers 1, 2, 3 and 4, respectively with a kernel size of 5×5. Finally, a dense layer with a size of 10 is used for the classification process. The number of epochs are 50. The classification neural network consists of a single hidden layer and an output layer, since we have 10 speakers to identify. In Table 5.1, the accuracy of the three proposed models is estimated. The accuracy reaches 98.75% for CNN of three layer. The training progress for the four models is shown in Figs. 5.3, 5.4, 5.5, and 5.6.

Table 5.1 Recognition rate of cancellable biometric system for different models

Model	**Recognition rate %** **Cancellable voiceprint**
Benchmark model	**94.37**
Two-layer CNN	**97.5**
Three-layer CNN	**98.75**
Four-layer CNN	**98.75**



Results
Validation accuracy: 94.38%
Training finished: Reached final iteration

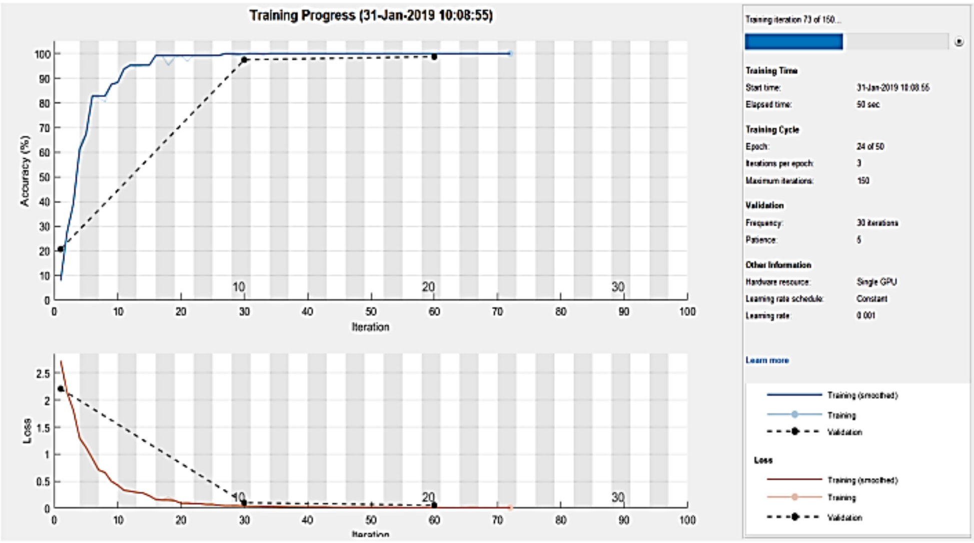
Training Time
Start time: 31-Jan-2019 10:15:53
Elapsed time: 1 min 33 sec

Training Cycle
Epoch: 50 of 50
Iteration: 150 of 150
Iterations per epoch: 3
Maximum iterations: 150

Validation
Frequency: 30 iterations
Patience: 5

Other Information
Hardware resource: Single GPU
Learning rate schedule: Constant
Learning rate: 0.001

Fig. 5.3 Training progress of the Benchmark model.



Training Progress (31-Jan-2019 10:08:55)

Training iteration 73 of 150...

Training Time

Start time: 31-Jan-2019 10:08:55

Elapsed time: 50 sec

Training Cycle

Epoch: 24 of 50

Iterations per epoch: 3

Maximum iterations: 150

Validation

Frequency: 30 iterations

Patience: 5

Other Information

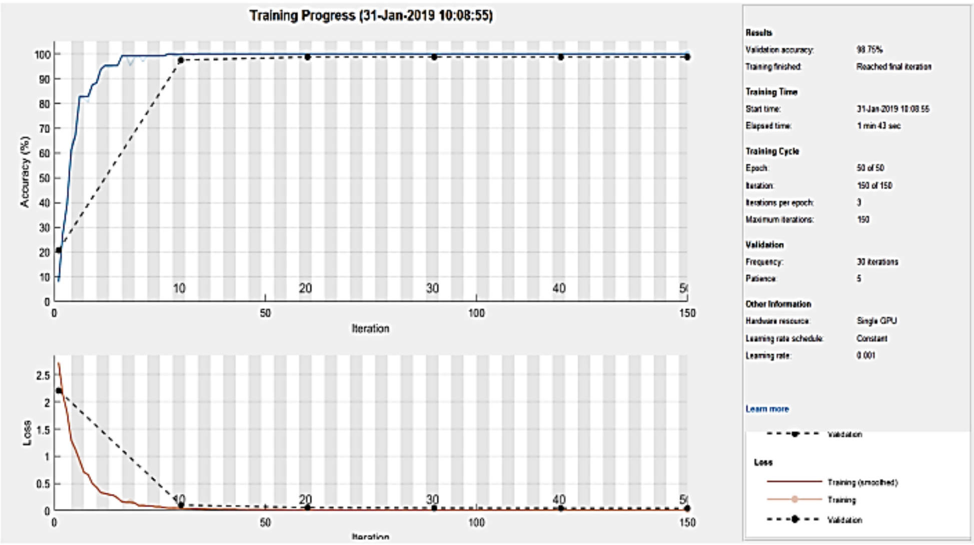
Hardware resource: Single GPU

Learning rate schedule: Constant

Learning rate: 0.001

Learn more

Fig. 5.4 Training progress of two-layer CNN model.



Training Progress (31-Jan-2019 10:08:55)

Results

Validation accuracy: 98.75%

Training finished: Reached final iteration

Training Time

Start time: 31-Jan-2019 10:08:55

Elapsed time: 1 min 43 sec

Training Cycle

Epoch: 50 of 50

Iteration: 150 of 150

Iterations per epoch: 3

Maximum iterations: 150

Validation

Frequency: 30 iterations

Patience: 5

Other Information

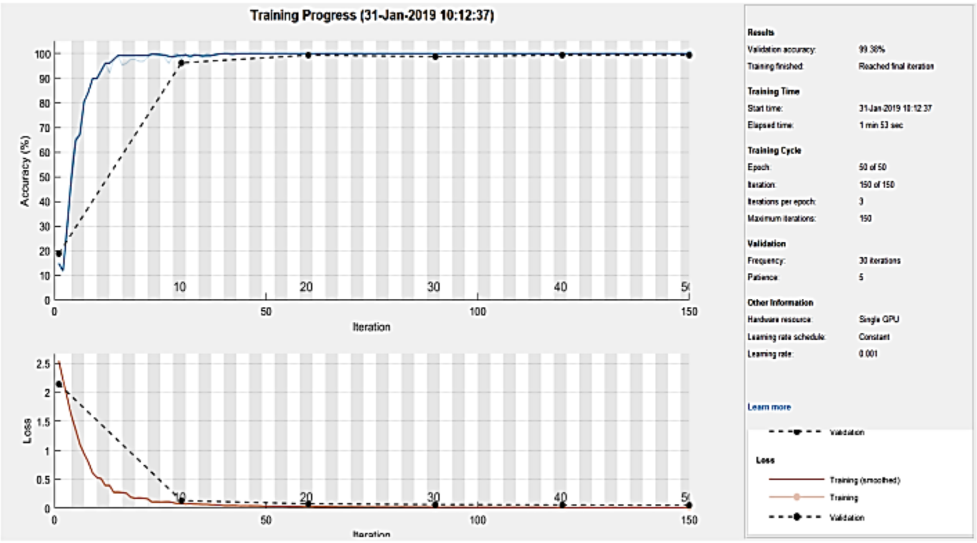
Hardware resource: Single GPU

Learning rate schedule: Constant

Learning rate: 0.001

Learn more

Fig. 5.5 Training progress of three-layer CNN model.



Results

Validation accuracy:	99.38%
Training finished:	Reached final iteration

Training Time

Start time:	31-Jan-2019 10:12:37
Elapsed time:	1 min 53 sec

Training Cycle

Epoch:	50 of 50
Iteration:	150 of 150
Iterations per epoch:	3
Maximum iterations:	150

Validation

Frequency:	30 iterations
Patience:	5

Other Information

Hardware resource:	Single GPU
Learning rate schedule:	Constant
Learning rate:	0.001

Fig. 5.6 Training progress of four-layer CNN model.

It is clear from Fig. 5.7 that the recognition rate is increased by increasing the CNN layers till 3 layers and then it becomes constant. The essence is that model 3 achieves the highest recognition rate.

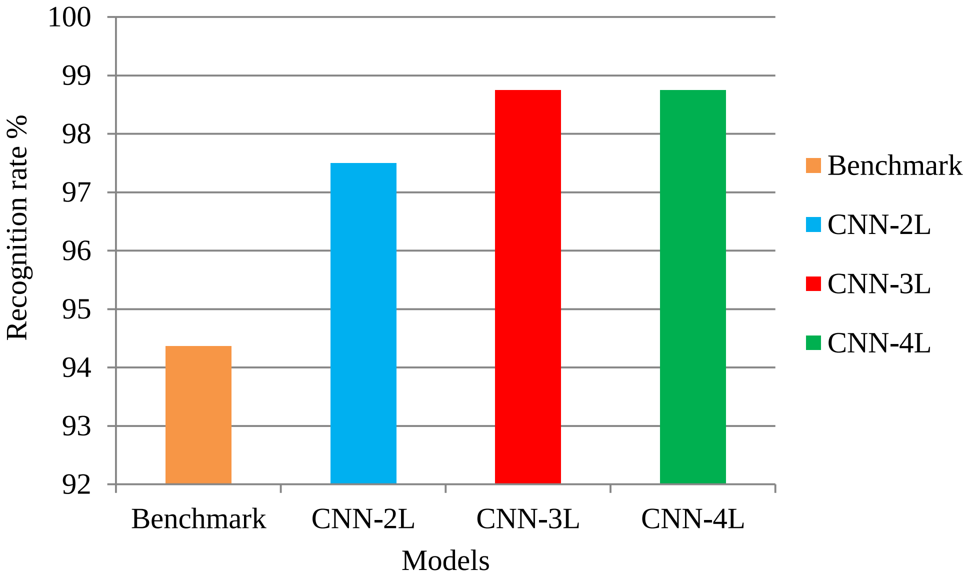


Fig. 5.7 Recognition rate versus number of CNN layers.

In this chapter, a proposed cancellable speaker recognition system was presented to achieve a high degree of security, and protect user's privacy. This system satisfies the required criteria as renewability, security and performance. It benefits from deep learning (CNN) by converting the speech signal into images suitable for the CNN to deal with. The achieved recognition rate that reaches 98.75% for three-layer CNN model proves high levels of detectability of speakers while keeping the speakers secure from any attempts of compromising. Three CNN models of different numbers of layers are used. A comparison was held between the three models and the conventional Benchmark model. The results reveal that the three models outperform the Benchmark model and the model of three-layer CNN has the best performance. The four-layer CNN model have the same performance of the three-layer CNN model.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

6.1 Conclusion

Speaker recognition is one of the most used biometric verification systems owing to its high importance in numerous applications of security and telecommunications. The main goal of an ASR system is to know who is speaking based on the voice characteristics. Many researches focus on text-dependent speaker recognition, which has a pre-knowledge of the speaker utterances. In this thesis, a text-independent speaker recognition system was considered, where no prior knowledge is accessible in the context of the speakers' utterances for all stages. This thesis provided a review of speaker identification system innovations. A speaker recognition system based on deep learning was suggested.

Most studies consider an ideal speaker recognition system without mismatching. Here, we considered a more realistic scenario. In real cases, some sort of mismatching exists such as noise, reverberation, and interference. This thesis presents an efficient approach for the classification of speech signals as reverberant or not. The reverberation is a severe effect encountered in closed rooms. So, it may affect subsequent processes and deteriorate speech processing system performance. The spectrograms are used as images generated from speech signals and are used with the CNN. The spectrograms and MFCCs are used as features with the LSTM-RNN. The two models have been presented and compared. Simulation results up to 100% classification accuracy have been obtained. This can help in performing an initial step in any speech processing system that comprises quality level classification.

Also, In this dissertation, a text-independent speaker recognition approach under mismatching has been implemented. The tools considered in this thesis are MFCCs, spectra, and log-spectra with the LSTM-RNN classifier. Also, speech enhancement was adopted with feature extraction methods to get the best performance, when utterances are degraded. The essence of the results shows that the system is capable of identifying speakers well, text independently, with the undistorted utterances and in the same recording situation with accuracy up to 98.7%. For distorted utterances, the performance degrades. A noticeable enhancement in the performance was achieved, especially when using spectrum features with spectral subtraction enhancement method with results up to 90%.

The Radon projections of the spectrograms of the speech signals have been used as features, since they are insensitive to noise and reverberation conditions. These features are extracted by applying the RT on the spectrogram patterns. The suggested approach computes the Radon projections for thirteen orientations and captures the acoustic characteristics of the spectrogram. The DCT has been applied on the Radon projections as it yields low-dimensional feature vectors. This technique is computationally efficient, text-independent, robust to variations and insensitive to additive noise.

The CNN-based feature extraction has been extended to the text-independent speaker recognition task. Also, the effect of reverberation on the speaker recognition has been addressed. All the speech signals are converted into images by obtaining their spectrograms. A proposed CNN model has been presented to enhance the performance of the system in the case of reverberant signals. The proposed model has been compared with the conventional Benchmark model. The performance of the ASR system has been measured by the recognition rate in the cases of clean and reverberant signals.

Most researches in speaker recognition applications consider an ideal scenario for training and testing without noise or interference. In real cases, some sources of interference may exist. In this thesis, we considered a realistic case of interference with the original signals. Hence, signal separation needs to be applied to reduce the interference effect. Different techniques of blind signal separation have been applied to improve the recognition rates of the text-independent speaker identification system. Blind signal separation is an important branch of signal processing, as it deals mainly with mixed signals, which are frequently encountered in real life. This thesis also covered the text-independent speaker recognition process assisted with signal separation and deep learning techniques. Two models of deep learning have been used: the LSTM-RNN and the deep CNN. Separation is performed in different domains and the results are compared to the conventional systems with clean and mixed utterances without separation. Results reveal that separation, especially in time domain, achieves higher recognition rates than those obtained in other domains. Classification rates of 97.33 and 96% have been obtained with three-layer CNN and LSTM-RNN.

The biometric patterns are non-revocable and vulnerable to confidentiality attacks. Cancellable biometrics has been presented to resolve these issues. In this thesis, we have presented an approach to generate cancellable templates from speech signals that can be replaceable and can be re-released, when compromised. The proposed approach allows the generation of variable speech templates without likability. Since the voice is the used biometric in the ASR system, it is important to save the patterns from being attacked, and keep the user's confidentiality. The proposed approach is based on converting the speech signals into images by obtaining their spectrograms to deal with using a deep CNN. Only, a patch of each spectrogram is used as an input to the network based on a user-specific key. The simulation results reveal that the suggested approach is practical and it satisfy the desired criteria such as renewability, security, and high performance. The accuracy of the proposed scheme reaches 98.75% with CNN of three layers.

6.2 Future work

The future work can be summarized in the forthcoming points:

- Performing signal separation in the presence of noise with deep learning.
- Performing signal separation in the presence of reverberation with deep learning.
- Cancellable speaker recognition with degradation models and deep learning.
- Demonstration of the effect of using a bi-directional LSTM instead of LSTM.
- Utilization of the auto-encoders to denoise corrupted speech signals.

REFERENCES

[1] E. Pagnin, and A. Mitrokotsa, “Privacy-Preserving Biometric Authentication: Challenges and Directions”, Security and Communication Networks, Vol. 2017, pp. 1-9, 2017.

[2] John H. L. Hansen, Taufiq Hasan, “Speaker Recognition by Machines and Humans: A Tutorial Review”, IEEE Signal processing magazine, 32(6), pp. 74-99, 2015, DOI: 10.1109/MSP.2015.2462851

[3] Tirumala, Sreenivas Sremath, et al. “Speaker Identification Features Extraction Methods: A Systematic Review”, Expert Systems with Applications, Vol. 90, pp. 250-271, DOI.org/10.1016/j.eswa.2017.08.015, 2017.

[4] M. Faundez-Zanuy, “Biometric Security Technology”, IEEE, Aerospace and Electronic Systems Magazine, Vol. 21, pp.15-26, June 2006.

[5] J. A. Markowitz, “Voice Biometrics”, Article in Communications of the ACM September, Vol. 43, No. 9, pp. 66-73, 10.1145/348941.348995,2000.

[6] S. Tripathi and S. Bhatnagar, “Speaker Recognition”, IEEE, 2012 Third International Conference on Computer and Communication Technology, Allahabad, India 2012, pp. 283-287, DOI: 10.1109/ICCCT.2012.64, 2012.

[7] Boujnah, Sana, et al., “3-Step Speaker Identification Approach in Degraded Conditions,” IEEE, 15th International Multi-Conference on Systems, Signals & Devices (SSD), Tunisia pp, pp. 1100-1103, DOI: 10.1109/SSD.2018.8570611, 2018.

[8] Md. Murad Hossain, B. Ahmed, and M. Asrafi, “A Real Time Speaker Identification using Artificial Neural Network”, IEEE, 10th International Conference on Computer and Information Technology, US, pp. 1-5, 2007.

[9] Rajeev Kumar et al., “Multilingual Speaker Recognition Using Neural Network”, In: Proceedings of the Frontiers of Research on Speech and Music (FRSM-2009), pp. 1-8, Dec. 2009.

[10] Tudor Barbu, “A Supervised Text-Independent Speaker Recognition Approach”, International Journal of Electronics and Communication Engineering, Vol. 1, No. 9, pp. 2726-2730, 2007.

[11] Richard D. Peacocke and Daryl H. Graf, “An Introduction to Speech and Speaker Recognition”, IEEE, Readings in Human–Computer Interaction. Morgan Kaufmann, 1995. pp. 546-553, 1995.

[12] Fathi E. Abd El-Samie, “Information Security for Automatic Speaker Identification”, Springer Briefs in Speech Technology, pp. 1-122, 2011.

[13] Y. Hioka, J.W. Tang, and J. Wan, “Effect of Adding Artificial Reverberation to Speech-Like Masking Sound”, Applied Acoustics, pp. 171-178, 2016.

[14] D. C. B. Chan, “Blind Signal Separation”, dissertation submitted to the university of Cambridge for PHD, Jan. 1997.

[15] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, IEEE, International Conference on Acoustics, Speech, and Signal Processing, US, pp. 4072-4075, 2002.

[16] A. A. Fisusi, T. K. Yesufu, “Speaker Recognition Systems: A Tutorial”, African Journal of Information and Communication Technology, Vol. 3, No. 2, pp. 42-52, June 2007.

[17] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, “A Survey on Automatic Speaker Recognition Systems”, Signal Processing and Multimedia. Springer, Berlin, Heidelberg, pp. 134–145, 2010.

[18] T. Kinnunen, H. Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors”, Speech Communication, Elsevier, pp. 12-40, 2010.

[19] R. Togneri and D. Pullella, “An Overview of Speaker Identification: Accuracy and Robustness Issues”, IEEE, Circuits and Systems Magazine, pp. 23-61, 2011.

[20] Z. Aslan, and M. Akin, "Performing Accurate Speaker Recognition by Use of SVM and Cepstral Features", The International Journal of Energy & Engineering Sciences (IJEES), pp. 16-25, 2018.

[21] S. Nilufar, N. Ray, M. K. Islam Molla, and K. Hirose, "Spectrogram Based Features Selection Using Multiple Kernel Learning for Speech/Music Discrimination", IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 501-504, 2012.

[22] P. Anand, A. Kumar Singh, S. Srivastava, and B. Lall, "Few Shot Speaker Recognition Using Deep Neural Networks", arXiv preprint arXiv:1904.08775, Computer Science, Engineering, Apr. 2019.

[23] H. B. Kekre, V. Kulkarni, P. Gaikar, and N. Gupta , "Speaker Identification using Spectrograms of Varying Frame Sizes", International Journal of Computer Applications, 50, No.20, pp. 27-33, July 2012.

[24] R. B. Diwate, S. J. Alaspurkar, "Study of Different Algorithms for Pattern Matching", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 3, pp. 615-620, March 2013.

[25] S. B. Dhonde, and S. M. Jagade, "Pattern-Matching for Speaker Verification: A Review", Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Springer, Odisha, India, pp. 19–25, 2014.

[26] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview", Neural Networks 61 (2015), pp. 85–117, 2015.

[27] GALLO, Crescenzio. "Artificial Neural Networks Tutorial", Encyclopedia of Information Science and Technology, Third Edition. IGI Global, pp. 6369-6378, 2015

[28] Geoffrey Hinton, Li Deng, Dong Yu, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition", IEEE Signal Processing Magazine, pp. 1-17, Nov. 2012.

[29] Phil Kim, "MATLAB Deep Learning", With Machine Learning, Neural Networks and Artificial Intelligence 130, Library of Congress Control Number: 2017944429, DOI 10.1007/978-1-4842-2845-6, 2017.

[30] T. N. Sainath, O. Vinyals, A. Senior, and Hasim Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, pp. 4580-4584, 2015.

[31] A. Graves, "Long Short-Term Memory", In Supervised Sequence Labelling with Recurrent Neural Networks, Springer-Verlag Berlin Heidelberg, SCI 385, pp. 37–45, 2012.

[32] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey", IEEE Transactions on Neural Networks and Learning Systems, pp. 2222-2232 , 2016.

[33] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition", arXiv:1402.1128v1, 2014.

[34] Z. C. Lipton, J. Berkowitz, C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning", arXiv:1506.00019v4 [cs.LG], Oct 2015.

[35] O. Abdel-Hamid, Abdel-rahman Mohamed, H. Jiang, Li Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 10, pp. 1533-1545, Oct. 2014.

[36] V. Dumoulin and F. Visin. "A Guide to Convolution Arithmetic for Deep Learning". ArXive-prints, March 2016.

[37] Saad Albawi , Tareq Abed Mohammed, and Saad AL-Zawi, "Understanding of a Convolutional Neural Network", IEEE, International Conference on Engineering and Technology (ICET) 2017, Antalya, Turkey, pp.1-6, 2017.

[38]Y. LuCun, and Y. Bengio, “Convolutional Networks for Images, Speech, and Time-Series”, The handbook of brain theory and neural networks, pp. 255–258, Oct. 1998.

[39] Yaming Wang, “Robust Text-independent Speaker Identification in a Time-varying Noisy Environment”, Journal of Software, Vol. 7, No. 9, Sept. 2012.

[40] Yanick Lukic, Carlo Vogt, Oliver Durr, Thilo Stadelmann, “Speaker Identification and Clustering Using Convolutional Neural Networks”, IEEE International Workshop on Machine Learning For Signal Processing, pp. 1–6, 2016.

[41] Nayana P.K. et al., “Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods”, 7th International Conference on Advances in Computing & Communications, ICACC-2017 India, pp.47-54, August 2017.

[42] Yanpei Shi, Qiang Huang and Thomas Hain, “Improving Noise Robustness In Speaker Recognition Using A Two-Stage Attention Model”, arXiv:1909.11200v2 [eess.AS], 2020.

[43] N. Upadhyay, and A.Karmakar, “Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study”, Procedia Computer Science 54, pp. 574 – 584, 2015.

[44] Luo Jun, and Zhiming He, “Spectral Subtraction Speech Enhancement Technology Based on Fast Noise Estimation”, IEEE, International Conference on Information Engineering and Computer Science, China, pp. 1-3, DOI: 10.1109/ICIECS.2009.5362621, 2009.

[45] Kaladharan N, “Speech Enhancement by Spectral Subtraction Method”, International Journal of Computer Applications (0975 – 8887), Vol. 96, No.13, pp. 45-48, June 2014.

[46] Mihov, Slavy G., Ratcho M. Ivanov, and Angel N. Popov. “Denoising Speech Signals by Wavelet Transform”, Annual Journal of Electronics 6 (2009), pp. 2-5, 2009.

[47] S. Bharkad, and M. Kokare, "Rotation-invariant Fingerprint Matching Using Radon and DCT", Springer Science and Business Media LLC in Sādhanā, Vol. 42, pp. 2025-2039, Nov. 2017.

[48] D. Joshi, M. Deo Upadhayay, and S. D.t Joshi, "Robust Language and Speaker Identification Using Image Processing Techniques Combined with PCA", International Conference on Signal Processing and Communication (Icsc) India, IEEE, pp. 213-218, 2013.

[49] P. K. Ajmera, D. V. Jadhav, and R. S. Holambe, "Text-independent Speaker Identification Using Radon and Discrete Cosine Transforms Based Features from Speech Spectrogram", Pattern Recognition, Elsevier, pp. 2749–2759, 2011.

[50] J. F. Cardoso, "Blind Signal Separation: Statistical Principles", Proceedings of the IEEE, Vol. 86, No. 10, pp. 2009-2025, Oct. 1998.

[51] Ricky Der, "Blind Signal Separation", Telecommunication and Signal Processing Laboratory, McGill University, Sep. 11, 2001.

[52] J. V. de Laar, E. A. P. Habets, J. D. P. A. Peters, and P. A. M. Lokkart, "Adaptive Blind Audio Signal Separation on a DSP", Computer Science, pp. 475-479, 2001.

[53] H. Hammam, A. E. Abu El-Azm, M. E. Elhalawany, and F. E. Abd El-Samie, "Simultaneous Blind Signal Separation and Denoising", 2008 International Conference on Computer Engineering & Systems, IEEE, Cairo, Egypt, pp. 107-112, 2008.

[54] B. Neyazi Badawi , A. Z. Mahmoud, H. Soliman Seddeq, El-Sayed M. El-Rabie, M. I. Dessouky, F. E. Abd El-Samie, "Sensitivity of Pitch Frequency Estimation to Reverberation Effect", Menoufia Journal of Electronic Engineering Research (MJEER), Vol. 29, Issue 1, pp. 50-55, Winter and Spring 2020.

[55] Y. Hioka, J. W. Tang, and J. Wan, **"**Effect of adding artificial reverberation to speech-like masking sound**",** Applied Acoustics, pp. 171–178, 2016.

[56] B. Cauchi, and H. Javed et al., “Perceptual and Instrumental Evaluation of the Perceived Level of Reverberation”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 629-633, 2016.

[57] J. Eaton, and P. A. Naylor, “Reverberation Time Estimation on The ACE Corpus Using The SDD Method”, ArXiv 2015, Computer Science, Oct. 2015.

[58] P. P. Parada, D. Sharma, P. A. Naylor, T. van Waterschoot, “Reverberant Speech Recognition: A Phoneme Analysis”, in Proc. 2014 IEEE Global Conference on Signal and Information Processing, Atlanta, Georgia, pp. 567-571, 2014.

[59] “ Magic Data Technology Co., Ltd., “http://WWW.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101”, 05/2019”.

[60] B. Choudhury, and P. Then,“ A Survey on Biometrics and Cancelable Biometrics Systems”, International Journal of Image and Graphics, Vol. 18, No. 1, pp. 1-39, 2018.

[61] W. Xu, Q. He, Y. Li, and T. Li,“ Cancelable Voiceprint Templates Based on Knowledge Signatures”, IEEE, International Symposium on Electronic Commerce and Security, pp. 412-415, 2008.

[62] W. Xu, M. Cheng,“ Cancelable Voiceprint Template Based on Chaff-Points-Mixture Method”, IEEE, International Conference on Computational Intelligence and Security, Suzhou, China, pp. 263-266, 2008.

[63] A. Mtibaa, D. Petrovska-Delacr´etaz, and A. Ben Hamida, “Cancelable Speaker Verification System Based on Binary Gaussian Mixtures”, IEEE, 4th International Conference on Advanced Technologies for Signal and Image Processing, ATSIP, sousse, Tunisia, 2018.

[64] E. Maiorana, P. Campisi, J. Fierrez, J. Ortega-Garcia, and A. Neri, “Cancelable Templates for Sequence-Based Biometrics with Application to On-line Signature Recognition”, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, Vol. 40, No. 3, pp. 525-538, May 2010.

ملخص الرسالة

يعتبر نظام التعرف على المتحدث من أهم الأنظمة المستخدمة في مجال معالجة الإشارات الكلامية حيث يستخدم في تطبيقات التحقق والأمان. وهو يعنى تحديد هوية المتحدث من خلال الكلام المنطوق، لذلك فهو نظام يحاكى الجهاز السمعي للإنسان فعندما يسمع صوت شخص يستطيع تحديد هويته ويتم ذلك عن طريق أخذ بصمه صوت لكل شخص. وتعتمد عملية التعرف أساسا على مرحلتين: مرحلة التدريب ومرحلة الاختبار وكلاهما يتضمن مرحلة تسمي استخلاص الخصائص، وهي تعني استخلاص بعض السمات المميزة لكل شخص والتي تميزه عن بقية الأشخاص وفي هذه المرحلة يتم تحويل الإشارة الصوتية الى مجموعة من المعاملات التي تُستخدم في عملية التدريب والاختبار. وتُستخدم هذه المعاملات أو الخصائص في تدريب واختبار خوارزميات مطابقة الأنماط وينقسم نظام التعرف على المتحدث الى نظام التعرف مستنداً على النص ونظام التعرف الغير مستند على نص، وتم التركيز في هذه الرسالة على النوع الثاني وهو نظام التعرف الغير مستند على نص وهو الأشمل حيث يتم التعرف على المتحدث دون التقييد بنص معين أو ثابت.

وقد تم استخدام أساليب التعلم العميق مع نظام التعرف على المتحدث مثل الشبكات العصبية التلافيفية و الشبكات العصبية ذات الذاكرة طويلة المدى. وتعتبر هذه التقنيات امتداد للشبكات العصبية ولكنها متعددة الطبقات، وتطبق الشبكات العصبية ذات الذاكرة طويلة المدى في التنبؤ بالنصوص وتصنيف السلاسل وتستخدم الشبكات العصبية التلافيفية لاستخلاص المميزات والتصنيف في نفس الوقت وهى شائعه الاستخدام في معالجة الصور وتصنيفها.

وحيث أن معظم الدراسات التي تتم في نظام التعرف على المتحدث تعمل على الحالة المثالية دون الأخذ في الاعتبار أي مصدر للتشويش، لذا تم الأخذ في الاعتبار بعض أنواع التشويش مثل الضوضاء، الصدى والتداخل ودراسة تأثير تلك العوامل على معدل التعرف على المتحدث واستخدام بعض أساليب التحسين للحصول على أداء أفضل لعملية التعرف على المتحدث. كما تم استخدام تحويل رادون في استخلاص الخصائص أيضا لتحسين أداء نظام التعرف على المتحدث في وجود الضوضاء. كما تقدم الرسالة بعض المقترحات لتحسين أداء نظام التعرف وزياده السرية باستخدام بصمة الصوت القابلة للإلغاء.

وقد تم تقسيم محتويات الرسالة إلى ستة فصول بيانها كالتالي:-

الفصل الاول:- يبدأ بمقدمة عن نظام التعرف على المتحدث وأنواعه وبعض تطبيقاته، وبعض التحديات التي تواجه وتؤثر على هذا النظام. وينتهى بتناول أهمية موضوع الرسالة والغرض منه وذكر مقدمة عن باقي الفصول.

الفصل الثاني:- وهو فصل مرجعي يناقش بعض الأساسيات عن الإشارة الصوتية وكيفيه تكوين الصوت في الإنسان وتحول الجهاز البشرى الى ما يكافئه هندسياً. كما يتناول أيضا تركيب نظام التعرف الأوتوماتيكي للمتحدث ويستعرض تقنيات استخلاص الخصائص ومنها معاملات ميل التردد والطيف المستخدمة في هذه الرسالة. كما يعرض أيضا بعض أنواع الشبكات المستخدمة في عمليه التصنيف. ويقدم أيضا بعض أساسيات التعلم العميق والنماذج المختلفة منه ويشرح النموذجيين المستخدمين وهم الشبكات العصبية التلافيفية والشبكات العصبية ذات الذاكرة طويلة المدى. ويعرض تكوين كل نموذج على حده. كما يعرض أيضا بعض أنظمة التعرف على المتحدث المستخدمة.

الفصل الثالث:- ويعرض النموذج الرياضي للطرق المستخدمة مثل الطرح الطيفي، تقنية تقليل الضوضاء باستخدام التحويل المويجي، تحويل رادون و تقنية فصل الإشارات.

الفصل الرابع:- ويعرض بعض التحديات التي تواجه نظام التعرف على المتحدث مثل الضوضاء والصدى والتداخل، وتأثير هذه العوامل على نظام التعرف على المتحدث. كما يقدم نبذه عن الضوضاء، ظاهره الصدى بالتفصيل وكيفية محاكاه الصدى وتأثيره على الإشارة الصوتية، وتداخل الاشارات الصوتية. كما يقدم أهميه طرق التحسين المستخدمة لمعالجة الإشارات الصوتية كمرحلة مبدئية في نظام التعرف و تحسين أدائه. كما يتم تقديم مقترح لتصنيف الإشارة الصوتية لمعرفة ما اذا كانت تحتوى على الصدى أم لا، وترجع أهمية معرفة ذلك لضرورة تقييم الإشارة لاستخدامها في تطبيقات أخرى. ويقدم هذا الفصل أيضا مقترح لنظام التعرف باستخدام التعلم العميق مع بعض طرق التحسين ودراسة تأثير الضوضاء والصدى على هذا النظام المقترح. ويتم تقديم مقترح أخر لتحسين أداء النظام المقترح وذلك باستخدام طريقة جديده لاستخلاص الخصائص باستخدام تحويل رادون. كما تم بناء نظام تعرف أخر معتمد على الشبكات العصبية التلافيفية ومن ثم دراسة تأثير الصدى على النظام المقترح وتغيير عدد طبقات تلك الشبكة لتحسين الأداء. كما يختص هذا الفصل بدراسة تأثير التداخل على النظام المقترح واستخدام تقنية فصل الإشارات لفصل الإشارات الصوتية المتداخلة وتحسين أداء نظام التعرف على المتحدث.

الفصل الخامس:- ويقدم مقترح لزياده خصوصية المستخدم وحماية البصمة الصوتية من الاختراق وذلك بإستخدام البصمة القابلة للإلغاء.

الفصل السادس:- يلخص النتائج التي حصلنا عليها ويعطى اتجاهات للعمل المستقبلي في هذا المجال.

وقد ذيلت الرسالة بقائمة المراجع.

جامعة المنوفية
كلية الهندسة الإلكترونية
قسم هندسة الإلكترونيات والاتصالات الكهربية

استخدام اساليب التعلم العميق لتحليل الاشارات الصوتية

رسالة مقدمـة للـحصول علـى درجـــة دكتور الفلسفة في العلوم الهندسية.
تخصص هندسة الالكترونيات والاتصالات الكهربية
مجال الرسالة: معالجة الاشارات
قسم هندسة الالكترونيات والاتصالات الكهربية

مقدمة مـــن

المهندسة/ سامية عبدالمنعم عمر قابل

بكالوريوس في الهندسة الإلكترونية - قسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعـــة المنوفية - جمهورية مصر العربية
ماجستير في العلوم الهندسية - قسم هندسة الالكترونيات و الاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعـــة المنوفية - جمهورية مصر العربية

لـجنة الإشراف

أ.د. / محمد محمد عبدالسلام نصار
أستاذ متفرغ بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

أ.د. / معوض ابراهيم دسوقي
أستاذ متفرغ بقسم هندسة الإلكترونيات والاتصالات الكهربية
كلية الهندسة الإلكترونية بمنوف - جامعة المنوفية

أ.د. / نبيل عبدالواحد اسماعيل
أستاذ متفرغ بقسم هندسة وعلوم الحاسب
كلية الهندسة الإلكترونية بمنوف - جامعة المنوفية

د. / عادل شاكر الفيشاوي
أستاذ متفرغ (أستاذ مساعد) بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

2020م

جامعة المنوفية
كلية الهندسة الإلكترونية
قسم هندسة الإلكترونيات والاتصالات الكهربية

استخدام اساليب التعلم العميق لتحليل الاشارات الصوتية

رسالة مقدمة للحصول على درجة دكتور الفلسفة في العلوم الهندسية.
تخصص هندسة الالكترونيات والاتصالات الكهربية
مجال الرسالة: معالجة الاشارات
قسم هندسة الالكترونيات والاتصالات الكهربية

مقدمة من

المهندسة/ سامية عبدالمنعم عمر قابل

بكالوريوس في الهندسة الإلكترونية - قسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية - جمهورية مصر العربية
ماجستير في العلوم الهندسية - قسم هندسة الالكترونيات و الاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية - جمهورية مصر العربية

لجنة الإشراف

أ.د. / محمد محمد عبدالسلام نصار (   )
أستاذ متفرغ بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

أ.د. / معوض ابراهيم دسوقي (   )
أستاذ متفرغ بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

أ.د. / نبيل عبدالواحد اسماعيل (   )
أستاذ متفرغ بقسم هندسة وعلوم الحاسب
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

د. / عادل شاكر الفيشاوي (   )
أستاذ متفرغ (أستاذ مساعد) بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية

2020م

جامعة المنوفية
كلية الهندسة الإلكترونية
قسم هندسة الإلكترونيات والاتصالات الكهربية

استخدام اساليب التعلم العميق لتحليل الاشارات الصوتية

رسالة مقدمة للحصول على درجة دكتور الفلسفة في العلوم الهندسية.
تخصص هندسة الالكترونيات والاتصالات الكهربية
مجال الرسالة: معالجة الاشارات
قسم هندسة الالكترونيات والاتصالات الكهربية

مقدمة من

المهندسة/ سامية عبدالمنعم عمر قابل

بكالوريوس في الهندسة الإلكترونية - قسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية - جمهورية مصر العربية
ماجستير في العلوم الهندسية - قسم هندسة الالكترونيات و الاتصالات الكهربية
كلية الهندسة الالكترونية بمنوف - جامعة المنوفية - جمهورية مصر العربية

لجنة الحكم و المناقشة

أ.د./ أشرف عبدالمنعم خلف ( )
رئيس قسم الهندسة الكهربية
كلية الهندسة - جامعة المنيا

أ.د./ معوض ابراهيم دسوقي ( )
أستاذ متفرغ بقسم هندسة الإلكترونيات والاتصالات الكهربية
كلية الهندسة الإلكترونية - جامعة المنوفية

أ.د. / نبيل عبدالواحد اسماعيل ( )
أستاذ متفرغ بقسم هندسة وعلوم الحاسب
كلية الهندسة الإلكترونية - جامعة المنوفية

أ.د./ أسامه فوزى زهران ( )
أستاذ بقسم هندسة الالكترونيات والاتصالات الكهربية
كلية الهندسة الالكترونية - جامعة المنوفية

2020م