UNIVERSITI PUTRA MALAYSIA

DIGITAL SPEECH WATERMARKING FOR ONLINE SPEAKER RECOGNITION SYSTEMS

MOHAMMAD ALI NEMATOLLAHI

DIGITAL SPEECH WATERMARKING FOR ONLINE SPEAKER RECOGNITION SYSTEMS

By

MOHAMMAD ALI NEMATOLLAHI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2015

DEDICATION

This thesis is dedicated to my father Abo-Al-Ghasem, my mother Ziba, my sisters Nazanin, Mahnaz, Saeedeh and Freshteh.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

DIGITAL SPEECH WATERMARKING FOR ONLINE SPEAKER RECOGNITION SYSTEMS

By

MOHAMMAD ALI NEMATOLLAHI

June 2015

Chairman: Syed Abdul Rahman Al-Haddad Bin Syed Mohamed, PhD
Faculty: Engineering

Speaker recognition is popular and feasible for online applications such as the telephone or network. However, low recognition performance and various vulnerable slots in online speaker recognition systems are two main problems. Although some of these slots can be secured by digital speech watermarking, applying robust watermark can still seriously degrade the recognition performance of online speaker recognition systems. The main aim of this thesis was to improve the security of the communication channel, robustness, and recognition performance of online speaker recognition systems by applying digital speech watermarking. In this thesis, Multi-Factor Authentication (MFA) method was used by a combination of PIN and voice biometric through the watermarks. For this reason, a double digital speech watermarking was developed to embed semi-fragile and robust watermarks simultaneously in the speech signal to provide tamper detection and proof of ownership respectively. For the blind semi-fragile digital speech watermarking technique, Discrete Wavelet Packet Transform (DWPT) and Quantization Index Modulation (QIM) were performed to embed the watermark in an angle of the wavelet's sub-bands where more speaker specific information was available. For watermarking the encrypted PIN in voice, a blind and robust digital speech watermarking was used by applying DWPT and multiplication. The PIN was embedded by manipulating the amplitude of the wavelet's sub-bands where less speaker specific information was available. A frame selection technique was also applied to weigh the amount of speaker-specific information available inside the speech frames. In the developed frame selection technique, Linear Predictive Analysis (LPA) was applied to separate the system features (formants) and source features (residual errors) of the speech frames. Then, a frequency weighted function was used to quantify the formants. High order correlation and high order statistics were used for weighting the residual errors. The lower frames' weight could be ignored for online speaker recognition systems but applied for digital speech watermarking.

TIMIT, MIT, and MOBIO speech corpuses were used for evaluating the developed systems. The experimental results showed that a combination of DWPT and multiplication for robust digital speech watermarking technique had higher robustness as compared to other robust

watermarking techniques, such as Discrete Wavelet Transform (DWT) with Singular Value Decomposition (SVD) and Lifting Wavelet Transform (LWT) with SVD, against different attacks such as filtering, additive noise, compression, re-quantization, resampling, and different signal processing attacks. Furthermore, this technique had less degradation on the performance of speaker recognition verification and identification which were 1.16% and 2.52% respectively. For semi-fragile watermark, the degradation for speaker verification and identification were 0.39 % and 0.97 % respectively which can be ignored. Twenty percent of the speech frames could be watermarked without serious degradation for the recognition performance of speaker recognition. The identification rate and Equal Error Rate (EER) were improved to 100% and 0% respectively by applying digital speech watermarking. As a conclusion, the digital speech watermarking can enhance the security of the online speaker recognition systems against spoofing and communication attacks while improving the recognition performance by solving problems and overcoming limitations.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia Sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

TERA AIR PERTUTURAN DIGITAL BAGI SISTEM PENGECAMAN SUARA DALAM TALIAN

Oleh

MOHAMMAD ALI NEMATOLLAHI

Jun 2015

Pengerusi: Syed Abdul Rahman Al-Haddad Bin Syed Mohamed, PhD
Fakulti: Kejuruteraan

Pengecaman suara adalah popular dan sesuai digunakan untuk aplikasi dalam talian seperti telefon atau rangkaian. Sungguhpun begitu, prestasi pengecaman suara yang rendah dan pelbagai slot ancaman di dalam bidang sistem pengecaman suara dalam talian merupakan dua masalah utama. Biarpun sebahagian daripada slot tersebut boleh dilindungi melalui tera air pertuturan digital, menggunakan tera air yang lasak telah menurunkan secara serius prestasi pengecaman sistem pengecaman suara dalam talian. Tujuan utama tesis ini adalah untuk memperbaiki saluran komunikasi dari segi keselamatan, kebolehtahanan dan prestasi pengecaman sistem pengecaman suara dalam talian dengan menggunakan tera air pertuturan digital. Di dalam tesis ini, kaedah Faktor-Pelbagai Pengesahan (MFA) telah digunakan dengan gabungan PIN dan biometrik suara melalui tera air tersebut. Untuk tujuan ini, sebuah tera air pertuturan digital secara berganda telah dibangunkan untuk menerapkan tera air separa-rapuh dan lasak secara langsung ke dalam isyarat suara bagi membolehkan sebarang gangguan pengesanan dan bukti pemilikan dapat diperolehi. Bagi teknik tera air pertuturan digital separa-rapuh yang rawak, Perubahan Paket Gelombang Kecil Diskret (DWPT) dan Modulasi Indeks Pengkuantuman (QIM) telah dilaksanakan untuk menerapkan tera air dalam sudut sub-pita gelombang kecil dimana lebih banyak suara khusus yang wujud. Bagi tera air melalui penyulitan PIN dalam suara, sebuah tera air pertuturan digital yang rawak dan lasak telah digunakan dengan mengaplikasikan DWPT dan pendaraban. PIN tersebut telah diterapkan dengan amplitud sub-pita gelombang kecil yang telah dimanipulasikan dimana tidak banyak suara khusus yang wujud. Tambahan pula, sebuah teknik pemilihan bingkai telah digunakan untuk mengukur kandungan maklumat suara-khusus yang wujud di dalam bingkai-bingkai pertututuran. Di dalam teknik pemilihan bingkai yang dibangunkan itu, LPA telah digunakan untuk memisahkan sifat-sifat maklumat (forman-forman) dan sifat-sifat sumber (ralat sisa) dalam bingkai-bingkai pertuturan tersebut. Kemudian, sebuah fungsi timbangan kekerapan telah digunakan untuk menyatakan kuantiti forman-forman. Korelasi tertib tinggi dan statistik-statistik tertib tinggi telah digunakan untuk mengukur ralat-ralat sisa tersebut. Ukuran bingkai yang rendah boleh diabaikan bagi sistem pengecaman suara dalam talian tetapi digunakan bagi tera air pertuturan digital.

Korpus pertuturan TIMIT, MIT dan MOBIO telah digunakan untuk menilai sistem pengecaman suara dalam talian tersebut. Keputusan eksperimen menunjukkan bahawa teknik gabungan DWPT dan pendaraban untuk tera air pertuturan digital yang lasak bukan sahaja mempunyai kebolehtahanan yang tinggi berbanding dengan teknik tera air lasak yang lain, seperti Perubahan Gelombang Kecil Diskret (DWT) dengan SVD dan Perubahan Gelombang Kecil Angkatan (LWT) dengan SVD, malahan terhadap serangan-serangan yang lain seperti penapisan, hingar tambahan, pemampatan, pengkuantuman-semula, persampelan semula, dan serangan-serangan pemprosesan isyarat yang lain. Tambahan pula, teknik ini mempunyai penurunan yang lebih kecil ke atas prestasi pengenalpastian dan pengesahan pengecaman suara dimana masing-masing adalah 1.16% dan 2.52%. Bagi tera air separa-rapuh, penurunan pengesahan suara dan pengenalpastian masing-masing adalah 0.39% dan 0.97% dimana ianya boleh diabaikan. Dua puluh peratus daripada bingkai pertuturan boleh diterapkan tera air tanpa penurunan secara serius kepada prestasi pengecaman suara. Kadar pengenalpastian dan Kadar Ralat Sama (EER) telah diperbaiki masing-masing kepada 100% dan 0% dengan menggunakan tera air pertuturan digital. Sebagai kesimpulan, tera air pertuturan digital mampu meningkatkan keselamatan sistem pengecaman suara secara dalam talian terhadap penipuan dan serangan-serangan komunikasi, prestasi pengecaman dan menyelesaikan banyak masalah serta kelemahan.

.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank and pray to God for blessing my postgraduate study and giving me strength and courage.
I also would like to thank my father and mother for their encouragement and support by praying for me to do my thesis successfully.
Furthermore, I would like to express my gratitude and convey my thanks to my supervisor, **A.P. Dr. S.A.R Al-Haddad Bin Syed Mohamed,** for his supervision, advice, and guidance from the early stage of this research until the completion of the research.
Similarly, I would like to record my gratitude to my co-supervisors **Prof. Dr. Iqbal M Saripan** and **A.P. Dr. Shyamala Doraisamy,** for their guidance, advice, and continuous support during the entire course of this research.
I would especially like to mention my sincere gratitude to the Head of Department of Computer and Communication System Engineering, **Dr. Ahmad Shukri Bin Muhammad Noor**, for his continuous support during my thesis work.
I also would like to express my deepest thanks to the faculty member of Tehran University, **Dr Mohammad Ali Akhaee,** for his guidance and support.
I am deeply thankful to **Madam Sharon Goh,** all of my dear classmates and friends, and others who have supported me continuously during the entire course of this research.
I also thank **Mr. Azlan** for providing a prefect laboratory environment with great support and assistance.
I would like to express my deepest regards and blessings to my family who have supported me continuously during my study until the completion of my thesis work.
Lastly, I would like to express my special thanks to University Putra Malaysia for providing a beautiful, peaceful, and calm academic environment for research and study.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Syed Abdul Rahman Al-Haddad Bin Syed Mohamed, PhD
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Iqbal M Saripan, PhD
Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Shyamala Doraisamy, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by Graduate Student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced; this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

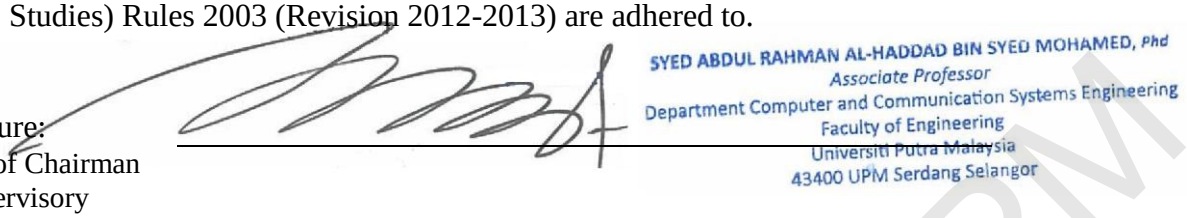
Signature: _____ Date: _____

Name and Matric No.: Mohammad Ali Nematollahi, GS31565

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.



Signature: _____
Name of Chairman of Supervisory Committee: Syed Abdul Rahman Al-Haddad Bin Syed Mohamed, PhD

Signature: _____
Name of Member of Supervisory Committee: Iqbal M Saripan, PhD

Signature: _____
Name of Member of Supervisory Committee: Shyamala Doraisamy, PhD

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

ACELP	Algebraic Code Excitation Linear Prediction
ADPCM	Adaptive Differential Pulse Code Modulation
ANN	Artificial Neural Network
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BPF	Band Pass Filter
bps	Bit per Second
BSC	Binary Symmetric Channel
CELP	Code Excitation Linear Prediction
CLT	Central Limit Theorem
CMS	Cepstral Mean Subtraction
DCF	Decision Cost Function
DCT	Discrete Cosine Transform
DET	Decision Error Trade-off
DFT	Discrete Fourier Transform
DMQ	Dither Modulation Quantization
DRT	Diagnostic Rhyme Test
DWPT	Discrete Wavelet Packet Transform
DWT	Discrete Wavelet Transform
EBU	European Broadcasting Union
EER	Equal Error Rate
EM	Expected Maximization
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FNR	False Negative Rate
FPR	False Positive Rate
FRR	False Rejection Rate
GGD	Generalized Gaussian Distribution
GLDS	Generalized Linear Discriminative Sequence
GMM	Gaussian Mixture Model
HAS	Human Auditory System
HMM	Hidden Markov Model
HOCOR	Haar Octave Coefficient of Residual
HOS	Higher Order Spectral (Statistics)
HVS	Human Vision System
IUT-I	International Telecommunications Union
JFA	Joint Factor Analysis
LAR	Log Area Ratio
LBG	Linde Buzo Gray
LPA	Linear Predictive Analysis
LPC	Linear Predictive Coefficient
LPCC	Linear Predictive Cepsstral Coefficient
LPF	Low Pass Filter
LSP	Line Spectrum Pair
LPRC	LP-Residual Cepsrum Coefficient
LVQ	Learning Vector Quantization

LWT	Lifting Wavelet Transform
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstrum Coefficient
MLLR	Maximum Likelihood Linear Regression
MNB	Measuring Normalized Blocks
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
MSE	Mean Squared Error
MSF	Magnitude Sum Function
NAP	Nuisance Attribute Project
PCA	Principle Component Analysis
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PIN	Personal Identification Number
PMC	Parallel Model Combination
PQ	Pre-Quantization
PSK	Probabilistic Sequence Kernel
PSNR	Peak Signal-to-Noise Ratio
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
QIM	Quantization Index Modulation
RASTA	RelAtive SpecTrAl
RDM	Rational Dither Modulation
RLP	Regularized Linear Prediction
RMSE	Root Mean Squared Error
SAD	Speech Activity Detection
SD	Spectral Distortion
SDG	Subjective Difference Grade
SEGSNR	Segmental Signal-to-Noise Ratio
SMS	Speaker Model Synthesis
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SQAM	Sound Quality Assessment Material
STE	Short Time Energy
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWLP	Stabilized Weighted Linear Prediction
SWR	Signal-to-Watermark Ratio
UBM	Universal Background Model
VAD	Voice Activity Detection
VoIP	Voice over IP
VQ	Vector Quantization
WCCN	Within-Class Covariance Normalization
WLP	Weighted Linear Prediction
XOR	Exclusive OR
ZCC	Zero Crossing Count

CHAPTER 1

INTRODUCTION

1.1 Overview

Many biometric features such as face, iris, hand, fingerprint, and voice have been used for biometric systems. However, each biometric technique has its pros and cons which are summarized in Table 1.1. Although the amount of False Non-Match Rate (FNMR) and False Match Rate (FMR) for the speech biometric technique is not accurate as compared to other biometric techniques, speech biometric is still suitable technique due to cost-effectiveness, less signal processing complexity (one-dimensional nature), and ease of use with less restrictions for recording.

Table 1.1. Comparison among different biometric recognition techniques (Motwani, 2010).

Attribute / Biometric	FNMR (%)	FMR (%)	Ease of use	Template size	Sensor cost	Long term stability	User acceptance	Variability
Face	4	10	M	1 kB	≈ 100$	M	H	Head pose, lighting, background, glasses, hair, facial expression, age.
Speech	15	3	H	2-3 kB	≈ 25 $	M	H	Illness, age, stress, fatigue, environment.
Iris	6	0.001	L	256 Byte	≈ 400 $	H	L	Poor lighting, eye position.
Fingerprint	2	2	M	0.5 kB	≈ 200$	H	M	Dryness, sensor noise, dirt, bruises.
Hand	1.5	1.5	H	0.1 kB	≈ 1500$	M	M	Hand injury, age.

- H=High, M=Medium, L=Low

Speech is the most important form of human communication as it reveals valuable information about a speaker. Speaker recognition is a kind of speech recognition system with speaker identification which involves identifying an unknown speaker by using a population of known speakers. This system also has speaker verification, as the most popular type of general biometric verification method which aims to verify the identity of a given speaker from a population of known speakers. Figure 1.1 shows how the technology of speaker recognition is a branch of speech processing science (Reynolds, 2010).

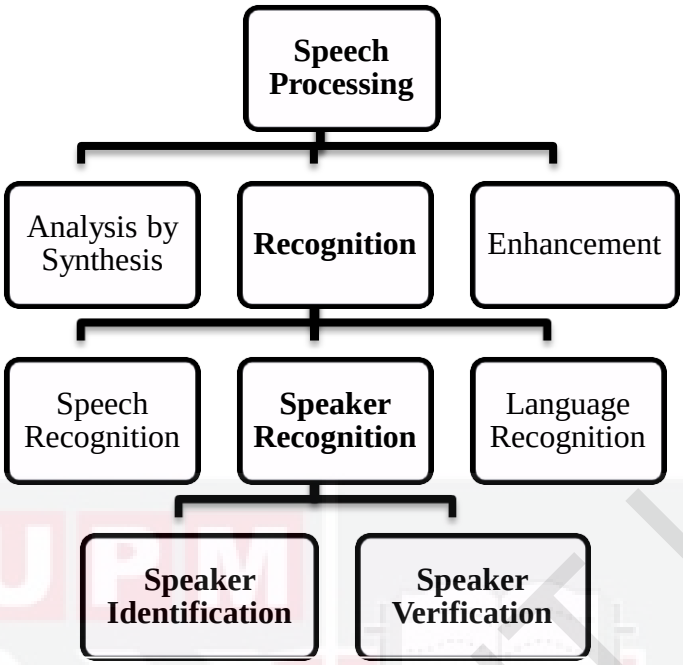


Figure 1.1. Taxonomy of speech processing (Reynolds, 2010)

Speaker recognition is designed as a system of pattern recognition. Firstly, a speech signal is sampled, quantized, and filtered. Then, it is used for the extraction of acoustic features. Secondly, this system uses acoustic features for training a speaker model. In the recognition (testing) phase, the extracted features from a test speech signal are matched with a speaker model for scoring. Figure 1.3 presents enrolment and recognition phases in an online recognition system which can be online speaker verification (with the result of acceptance or rejection) or online speaker identification (with the result of speaker ID).

Speaker recognition has applications for forensics, structuring audio information, games, access control and secured transition on telecommunication, and telephone banking as in Figure 1.2. The demand for speaker recognition applications comes from various fields, namely, tele-commerce, automobile industry, robotics, forensics, airports, smart homes, office environments, and law courts. The speaker recognition system may not be popular for on-site application (where the person needs to be in front of the system to be recognized) due to its inability to provide a certain level of reliability and security as compared to other biometric recognition techniques such as iris printing and fingerprinting. However, this system is still popular for online applications where the person can access the system through a remote terminal such as telephone or network (Bimbot et al., 2004). Online speaker recognition is also feasible for biometric system developers due to three main reasons (Fazel and Chakrabartty, 2011). Firstly, speech is easy to be produced, captured, and transmitted as it has a lower cost compared to other biometric recognition techniques. Secondly, speech is non-invasive and does not need direct contact with or to be perceived by an individual. Thirdly, speech can reveal information about an individual's gender, age, and emotion which is hidden from other biometric recognition techniques such as iris printing and fingerprinting (Fazel and Chakrabartty, 2011). Traditionally, the main application for speaker recognition is to be useful wherever there is a need to recognize a speaker's conversation over voice

channel such as telephone, wireless phone, or Voice over IP (VoIP). Secure and robust recognition is the primary concern of an online speaker recognition system for commercial applications. In this thesis, the main interest is increasing the security of communication channel and recognition performance of speaker recognition systems.

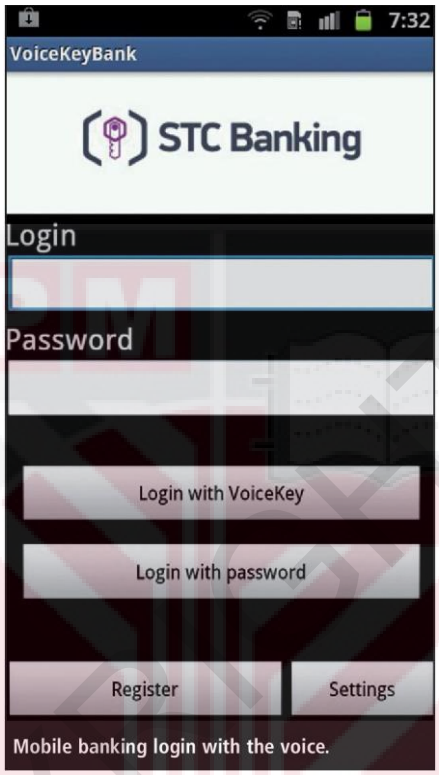


Figure 1.2. Android apps of voicekey for smartphone (Khitrov, 2013)

Online speaker recognition systems must have enough security and robustness to operate in real-word environments. However, there are potential vulnerable cracks which threaten online speaker recognition systems. In (Fatindez-Zanuy, 2004), eight points of vulnerability in these types of online biometric systems are discussed. Figure 1.3 shows that: (a) In a speech recorder (a microphone or sensor), an impostor may try to trick the system by imitating the speech of an authorized user; (b) In the transmission channel, the impostor, by locating in a different origin, may influence the system's security by using a stored tape of the authorize user, especially when online communication takes place on a wireless channel; (c) The feature extractor may be cracked by forcing it to produce the impostor's desired features; (d) The originally extracted features can be attacked by transmitting them to a classifier especially for online recognition purposes. Some applications may try to extract the features and send them to an online host for recognition as less amount of information is needed for transmission; (e) Other attacks may happen by manipulating the scores which are produced by the classifier, regardless of the features' sets; (f) The database can be altered and hence degrade the performance of an online speaker recognition system; (g) During transmission and storage a template which is used to extract the features, the impostor can replace these features easily; (h) Efforts to bypass the whole online speaker recognition

system can be done by manipulating the decision.

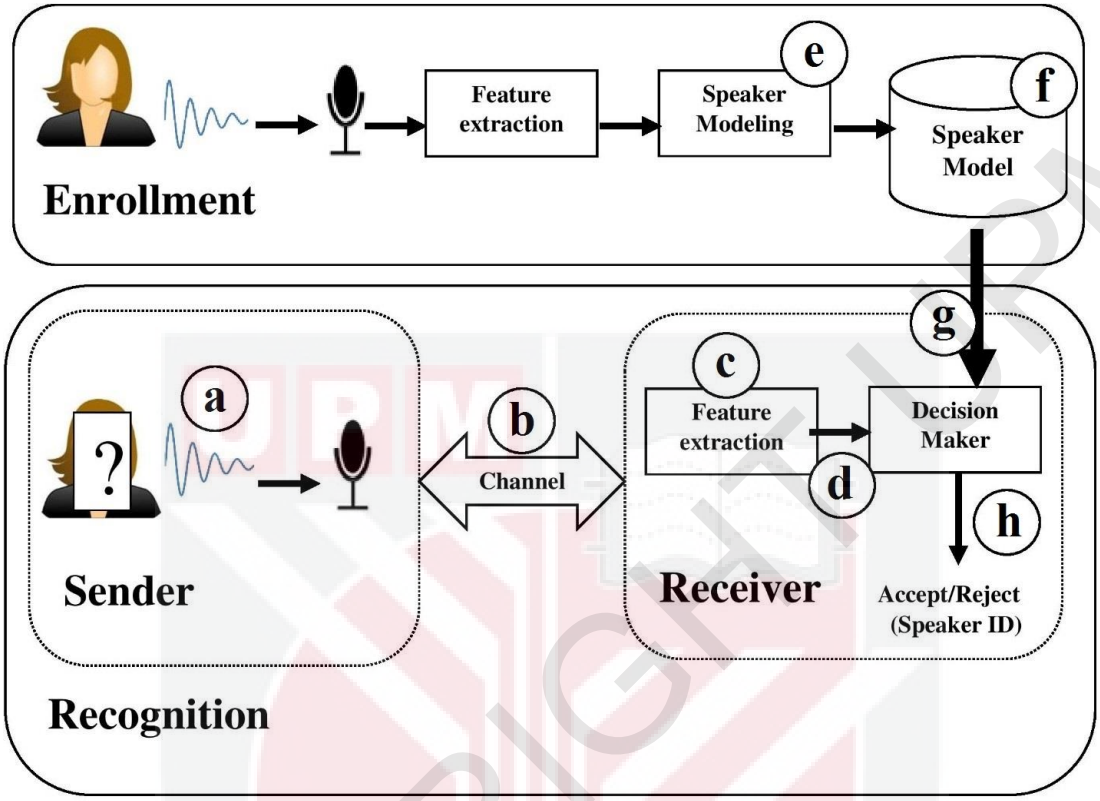


Figure 1.3. Eight vulnerable slots are presented in online speaker recognition systems (Fatindez-Zanuy, 2004; Wu et al., 2015)

All these attacks can be divided into two main classes (Fatindez-Zanuy, 2004). Firstly, communication attacks or "replay attacks" happen (b, d, g, and h) when the impostor tries to capture the speaker's feature. These attacks are due to an unsecured transmission channel. These types of attacks can be protected and secured by time stamp and watermarking. Recently, speech watermarking is used to secure the communication channel against intentional and unintentional attacks for speaker verification and identification purpose (Faundez-Zanuy et al., 2006; Al-Nuaimy et al., 2011; Faundez-Zanuy et al., 2007). Secondly, "system module attacks" (a, c, e, and f) happen when the impostor tries to break the module by generating huge random data. For instance, in the case of online speaker recognition, the impostor may use tremendous efforts to synthesize the speech, the features or the template of an authorized user. Although these types of attack can be secured by prohibiting extra access for one account in a special time period, other accounts can still be faked.

As discussed, speaker recognition technology has many benefits; however, two main problems still exist. Firstly, the recognition performance of online speaker recognition is not high enough (as it cannot provide 100% correct recognition). The user is inconvenienced by having false nonmatching (Khitrov, 2013; O'Gorman, 2003). Secondly, online speaker recognition is widely used in unattended telephony applications which can be exposed to

malicious manipulation and interference compared to other biometrics. Therefore, speaker recognition has more potential for spoofing attacks. According to a recent study (Wu et al., 2015), spoofing attack is the most potentially possible attack for an online speaker recognition system. Such attack can be divided into four attack classes such as impersonation, replay, speech synthesis, and voice conversion. Table 1.2 presents the most recent studies on the performance of the online speaker recognition system under various spoofing attacks. Table 1.2 is based on a recent survey paper (Wu et al., 2015). In addition to this vulnerability risk, high False Acceptance Rate (FAR) also can threaten the security of online speaker recognition systems. For this reason, today's technology has applied multimodal and Multi Factor Authentication (MFA) to solve the recognition performance and security problems by combining different authentication factors.

Table 1.2. Different studies on effect of various spoofing attack on speaker recognition performance.

Spoofing Attack	Practicality	Vulnerability	Countermeasure	Study	EER/FAR (%) before spoofing attack	EER/FAR (%) after spoofing attack
Impersonation	Low	Low	Non-existent	(Hautamäki et al., 2013)	9.03	11.61
Replay	High	High	Low	(Villalba and Lleida, 2011)	0.71	20
Speech synthesis	Medium to high	High	Medium	(De Leon, 2012)	0.28 0.00	86 81
Voice conversion	Medium to high	High	Medium	(Kinnunen et al., 2012)	3.24	7.61
				(Wu, 2012)	2.99	11.18
				(Alegre et al., 2012)	4.80	64.30
				(Alegre et al., 2013)	5.60	24.40
				(Alegre et al., 2013)	3.03	20.2
				(Kons and Aronowitz, 2013)	1.00	2.90

1.2 Problem Statements

Watermarking is applied as a possible means to enhance the security of speaker recognition systems against communication and spoofing attacks. (Faundez-Zanuy et al., 2006; Al-Nuaimy et al., 2011; Faundez-Zanuy et al., 2007). For this reason, the watermark is embedded to verify the authenticity of the transmitter (i.e. sensor and feature extractors), and the integrity of the entire authentication mechanism. Furthermore, multimodal systems use biometric watermarking to embed one modality into a second modality as a whole to improve recognition performance (Bartlow et al., 2007; Noore et al., 2007). However, multimodal systems only improve the recognition performance but it cannot improve the security due to using the same factor which is suffering from similar vulnerabilities. The majority of the

multimodal systems are unrealistic due to high cost and using different sensor devices (Huber et al., 2011). Therefore, a few MFA systems have proposed to apply another authentication factor (i.e. PIN or token) for embedding into the biometric signal for increased recognition performance and security of the overall system (Huber et al., 2011; Jain and Uludag, 2003; Rajibul Islam et al., 2008).
However, applying robust watermarking can seriously degrade the recognition performance. Since the main aim of multimodal and MFA technologies are to enhance recognition performance, applying watermark technology in this context is questionable due to its potential degradation on recognition performance. This impact is caused by three main problems:

First, degradation of recognition performance due to robust watermarking happens when white pseudo-noise signal is added to each frame of the speech signal uniformly. However, the amount of speaker-specific information is not uniformly distributed in the speech frames (Hyon, 2012; Lu and Dang, 2008). Watermarking all the speech frames can degrade the recognition performance of speaker recognition systems. Therefore, an investigation is needed to select the frames with less speaker-specific information for watermarking.

Second, available speech watermarking techniques (Faundez-Zanuy et al., 2010; Hofbauer et al., 2009) embed the watermark in the special frequency range or the speech formants. However, these techniques can seriously degrade the speaker recognition performance. Furthermore, the available digital watermarking techniques cannot satisfy the trade-off between different factors such as robustness, blindness, capacity, and imperceptibility. For example, blindness (Al-Nuaimy et al., 2011), robustness against various digital channel attacks such as compression, resampling, re-quantization, and filtering (Hagmüller et al., 2004; Hofbauer et al., 2005; Hofbauer et al., 2009), and high complexity in terms of time and memory (Bhat et al., 2011; Hu et al., 2014; Lei et al., 2012; Li and Kim, 2014; Yong-mei et al., 2013) are the main problems of the available digital watermarking techniques.

Third, watermarking and speaker recognition systems have opposite goals whenever the Signal-to-Watermark-Ratio (SWR) is decreased and the robustness of the watermark is increased. However, the speaker identification and verification performance can be decreased (Faundez-Zanuy et al., 2006; Al-Nuaimy et al., 2011; Baroughi and Craver, 2014; Faundez-Zanuy et al., 2007). Therefore, some researchers apply semi-fragile watermarking to reduce this impact on recognition performance (Hämmerle-Uhl et al., 2011). Although semi-fragile watermarking techniques can be used for tamper detection, a requirement is still needed for robust watermarking techniques to protect the ownership (Hämmerle-Uhl et al., 2011). Therefore, providing robustness and fragility at the same time for an online speaker recognition system by developing an MFA system is the main problem. Another factor such as blindness is also important for developing MFA based on digital speech watermarking.

1.3 Objectives

The main aim of this thesis is to improve the communication channel security and recognition performance of online speaker recognition systems through digital speech watermarking. For

this reason, three problems which are discussed in Section 0, need to be solved as following objectives:

1- To investigate and develop a frame selection technique to embed frames which can preserve the most discriminative speaker-specific features when watermarking to ensure that the recognition performance of the online speaker recognition system is not significantly compromised.
2- To design a blind double watermarking system to embed the semi-fragile and robust watermarks in more and less speaker specific frequency regions respectively.
3- To develop an MFA system based on a combination of PIN and voice biometric through digital speech watermarking to embed the speaker's PIN for improving the recognition performance and communication channel security.

1.4 Thesis Scope

This thesis analyzes online speaker recognition systems based on digital speech watermarking to improve the recognition performance and communication channel security. Such a system must have minimum degradation on the performance of speaker recognition, maximum communication channel security, and robustness against channel attacks. To achieve this aim, a new frame selection technique was developed based on speaker specific discrimination ability to select the speech frames with less speaker specific information for watermarking. In frame selection, speaker specific discrimination ability of the system (vocal tract) and source (glottal excitation) features were investigated. Furthermore, a digital speech watermarking was presented to provide robustness, capacity, and imperceptivity for this aim. Online speaker recognition based on digital speech watermarking has the following advantages over conventional speaker recognition systems:

1. Enhance the communication channel security and robustness of the online speaker recognition system against intentional and unintentional digital channel attacks.
2. Has the possibility of embedding biometric information, e.g., Eigen-face or PIN (Jain et al., 2006) as watermark data. The speech signal also always carries the watermark and any attempt to remove the watermark causes the signal to become worthless.
3. Use the watermark as a verification and authentication technique with efficient time and complexity for online speaker recognition systems (Blackledge and Farooq, 2008).

In this thesis, the feasibility of applying digital speech watermarking is investigated for two online speaker recognition systems: speaker verification and speaker identification. To investigate the feasibility of digital speech watermarking for online speaker recognition systems, two evaluation approaches were adopted. In the first approach, the effects of watermarking on the recognition performance of conventional speaker recognition systems was studied. In the second approach, a criterion based on bit error rate (BER) was established to evaluate the recognition performance of the online speaker recognition systems based on digital speech watermarking. This thesis only concentrated on wideband speech signals with 16 kHz sampling rate which were acquired from TIMIT (Garofolo and Consortium, 1993), MIT, and MOBIO speech corpuses. For baseline speaker recognition, Mel Frequency Cepstrum Coefficients (MFCC), two baseline speaker verification systems with GMM-UBM (Reynolds et al., 2000), i-vector PLDA (Dehak et al., 2011; Kenny, 2012), and GMM speaker identification (Pathak and Raj, 2013; Reynolds, 1995) systems were implemented to study the effects of digital speech watermarking on the performance of the speaker recognition systems.

In this thesis, the watermarked speech signal was assumed to be always synchronized. The size of the frames, quantization parameters, watermark intensity, and threshold value were known at the receiver. Figure 1.4 presents the focus of this thesis to narrow down the constraints in the field of data hiding:

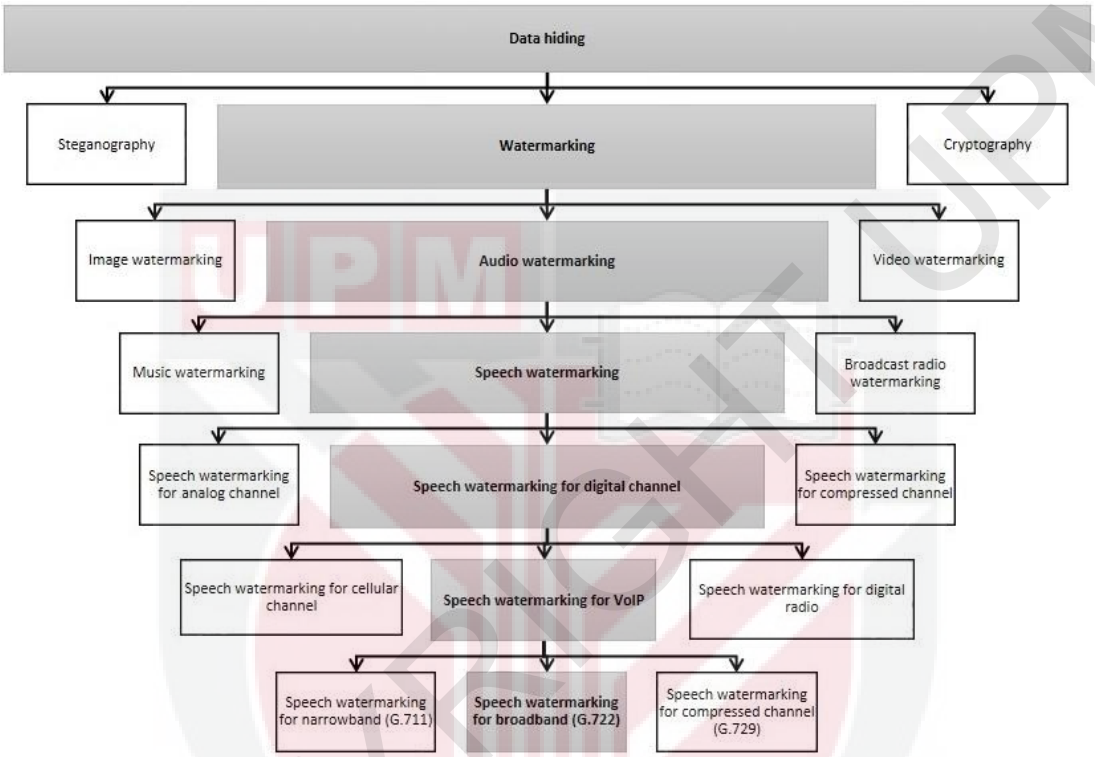


Figure 1.4. Thesis focus in the field of data hiding

1.5 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 of this thesis provides an overview on speech production system, speaker recognition technique and digital speech watermarking techniques. Firstly, basic speech processing information such as speech production model and speech characteristics is discussed. Secondly, a brief exploration on online speaker recognition techniques, problems, and solutions are explained. Furthermore, this section discusses the main parts in online speaker recognition systems such as pre-processing, robust feature extraction, robust speaker modelling, decision making, and different metrics for evaluation. Lastly, a basic theory, related, and pervious works in digital speech watermarking techniques are explained.

Chapter 3 presents an MFA approach based on digital speech watermarking for online speaker recognition with more details in four subsections. The first subsection discusses the overall framework of the proposed MFA system based on online speaker recognition and digital speech watermarking. The second subsection describes the proposed frame selection technique and how it can be applied in online speaker recognition systems properly. The third subsection presents a blind double digital speech watermarking technique. The fourth subsection discusses several advantages of the developed MFA. Furthermore, threat model and attack analysis of the developed MFA are evaluated.
Chapter 4 provides the simulation results and discussion on the proposed techniques which are presented in Chapter 3. In Chapter 4, firstly, the developed digital speech watermarking is compared to state-of-the-art digital watermarking techniques in terms of robustness, imperceptibility, capacity time, and memory. Secondly, the effects of digital speech watermarking on speaker verification and identification performance are discussed in two subsections respectively. Thirdly, the results and discussion on the developed frame selection technique is presented. Fourthly, the results of the developed MFA systems are presented. Finally, an overall discussion on the results are included in this chapter.

Lastly, Chapter 5 summarizes the contributions of this thesis, provides conclusion for this thesis, and suggests future works. The appendices provide some discussion about GGD shape estimation based on statistical moment of signal. Furthermore, the density computation of a ratio for two independent and normal variables is calculated. In addition, some discussion on tradeoff among various watermark criteria are presented by using many triangles. Furthermore, additional MATLAB script files which were implemented for the simulation of the results are provided.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter discusses the methodology for online speaker recognition systems based on digital speech watermarking. This chapter is organized as follow: firstly, the overall framework is discussed. Secondly, the investigation on speaker specific information in source and system features are discussed. Then, a speaker frame selection technique is developed based on system and source features. Thirdly, a blind double digital speech watermarking technique is presented. Fourthly, online speaker verification and identification systems are evaluated by using digital speech watermarking and speaker frame selection techniques. The systems can be considered as MFA systems based on speaker recognition and speech watermarking. Fifthly, threat model and attack analysis of the proposed MFA system are studied. Finally, test-bed environment for validation and simulation is discussed.
Figure 3.1 shows the conventional layered architecture for applying cryptography and watermarking. As seen, cryptographic algorithm is used on data to improve the security of watermark during channel transmission. This architecture can provide multiple lines of defense against malicious attacks and always superior to what a single line can do.

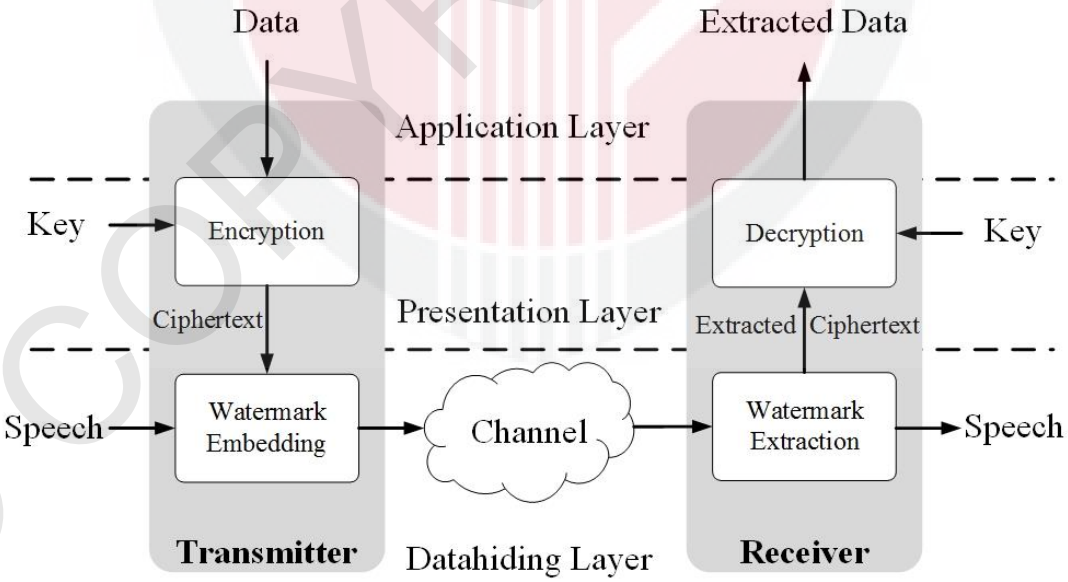


Figure 3.1. Different layers in designing a watermark system (Cox et al., 2006)

3.2 Overall Framework of MFA System Based on Online Speaker Recognition and Digital Speech Watermarking

This section reveals the overall framework of an MFA method based on online speaker recognition and digital speech watermarking to improve the security (in terms of communication channel attack, replay attack, and spoofing attack) and recognition performance. Different steps of the overall framework are shown in Figure 3.2. In the first step on the transmitter's side, it is important to make sure that the speech signal has not already been watermarked so that it can be analyzed by steganoanalysis techniques. Then in the second step, frame selection is used to select the less contributing frames for watermarking while the more suitable ones are kept for feature extraction in the speaker recognition system. The last step on the transmitter's side is to embed a unique speaker signature as the watermark in the less contributing frames. Before embedding, this signature is encrypted by a hashing function to provide enough level of security.

After passing through the communication channel, many factors such as attacks, noise, and environmental disturbances may corrupt the speech signal. On the receiver's side, after synchronization (which is beyond the scope of this thesis), the watermark is extracted from the speech signal. The speaker feature also is extracted for speaker recognition systems. The combination of speaker feature as a voice biometric and the decrypted watermark as a PIN can be considered as an MFA system. Applying this MFA method based on digital speech watermarking can improve the recognition performance and communication channel security of online speaker recognition systems. Each step in Figure 3.2 is discussed in detail in the following subsections.

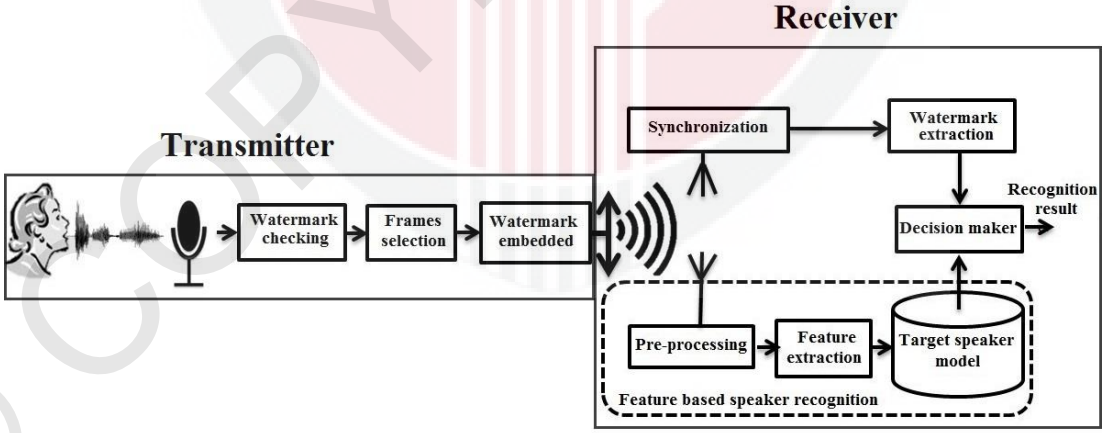


Figure 3.2. The developed framework based on speaker recognition and speech watermarking in parallel ways

3.2.1 Watermark Checking

In this step, firstly, the speech signal is sampled, quantized, and filtered. Secondly, the speech signal is checked for the existing watermarks. As mentioned earlier, a useful property of the watermark is that it cannot be destroyed without any serious degradation of the signal quality. If the watermark exists, it means some attacks (replay attacks) have happened. If the speech signal is clear from any watermark which can be checked by steganoanalysis techniques, it goes to the next step.

3.2.2 Proposed Frame Selection Technique

The proposed frame selection technique applied speaker discrimination information in source and system features to weigh the frames of the speech signal. For this reason, LPA was done for each frame to extract formants, gain, and residual errors. As discussed in Section 2.2, LPA can separate the source and system features of the speech signal. Therefore, LPA models the parameters of the vocal tract system; hence, glottal excitation remains in LP residual.

The frames were weighed in such a way that a higher value for the frame's weight could show better speaker discrimination. The frame's selection technique must be fast to ensure feasibility. The first speech signal was segmented into frames. Then some windowing functions like Ham and Hann were used. Next, pre-emphasis filter was used for each frame to remove the effect of lip radiation. LPA was then computed into LPCs, gain and LP residual. As mentioned in Section 2.2, LPCs were converted to formant bandwidth, formant frequency, and formant amplitude. Based on Equation (3.1), each frequency formant is weighted:

$$\text{Formant}=\frac{(\text{Formant amplitude})}{(\text{Formant bandwidth})} \quad (3.1)$$

Equation (3.1) finds the most predominant formants. It means whenever the amplitude is increased and the bandwidth is decreased, the sharper formant with more ability to discriminant the speaker is achieved. As mentioned in Section 2.2, the amplitude for the predominant formats is big (nominator) because the nearest pole to unit circle in Z-domain is a good formant candidate, with a bandwidth that is small (denominator). The speaker's discrimination weight is multiplied by these formants and the sum is computed. As a result, the weighted curve amplifies the formats which are located in the frequency area showing more speaker discrimination while suppressing the formants which are located in the lesser speaker discrimination. High order correlation, High Order Statistics (Spectral) (HOS) and energy of LP residual were also estimated. Although Gain (G) (which is the estimator of noise variance) is good in a clean environment, HOS of LP residual can still work appropriately in noisy conditions. All these weights were used to find the overall frames' weights. The process is illustrated in Figure 3.3:

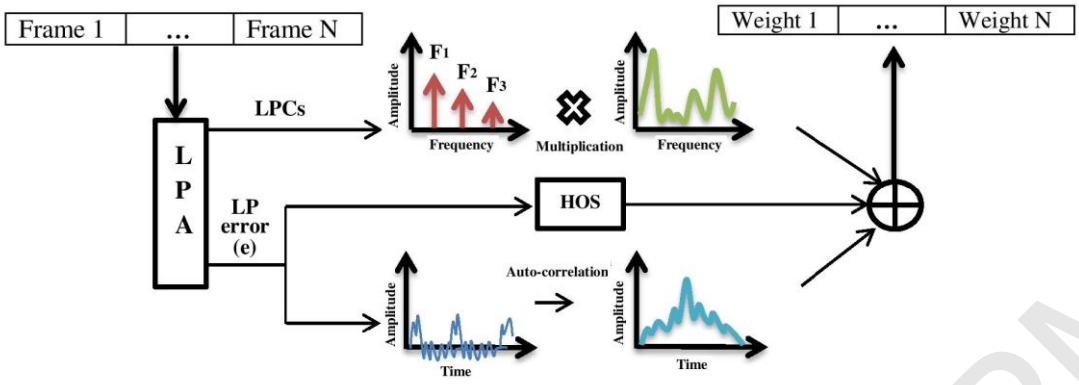


Figure 3.3 . The developed frame selection technique

By applying the proposed frame selection technique to N input frames, N weights were produced. The more weight for i_{th} shows more speaker discrimination ability of i_{th} frame. Depending on the time, memory, performance, cost and accuracy, the lower frames' weight can be ignored for speaker recognition or applied for digital speech watermarking. As a result, watermarking of the frames having lower weights can result in minimum degradation on the performance of the speaker recognition.

3.2.3 Proposed Double Digital Speech Watermarking Technique

In this part, the overall flow of double digital speech watermarking is discussed. Robust and semi-fragile watermarking have different security applications. Applying both of them simultaneously can improve the security of the proposed MFA system. Figure 3.4 shows the proposed double digital speech watermarking technique for online speaker recognition. As seen, OTP is embedded by using semi-fragile speech watermarking technique in sub-bands of wavelet where higher speaker specific information is available. The semi-fragile watermark is tied intrinsically to speaker biometric for tamper detection, and any attempt of adversary can destroy the semi-fragile watermark. Furthermore, the semi-fragile digital speech watermark technique has negligible degradation on recognition performance due to very small watermark intensity. At the same time, robust watermark is embedded into the rest of wavelet sub-bands, where less speaker specific information is available, to prevent interference between both robust and semi-fragile watermarking techniques. The robust watermark can protect the ownership and carries the PIN. Before the watermarks are embedded into the speech signal, Exclusive OR (XOR) is operated between key bits and watermark bits to improve the security of the watermark bits during channel transmission.

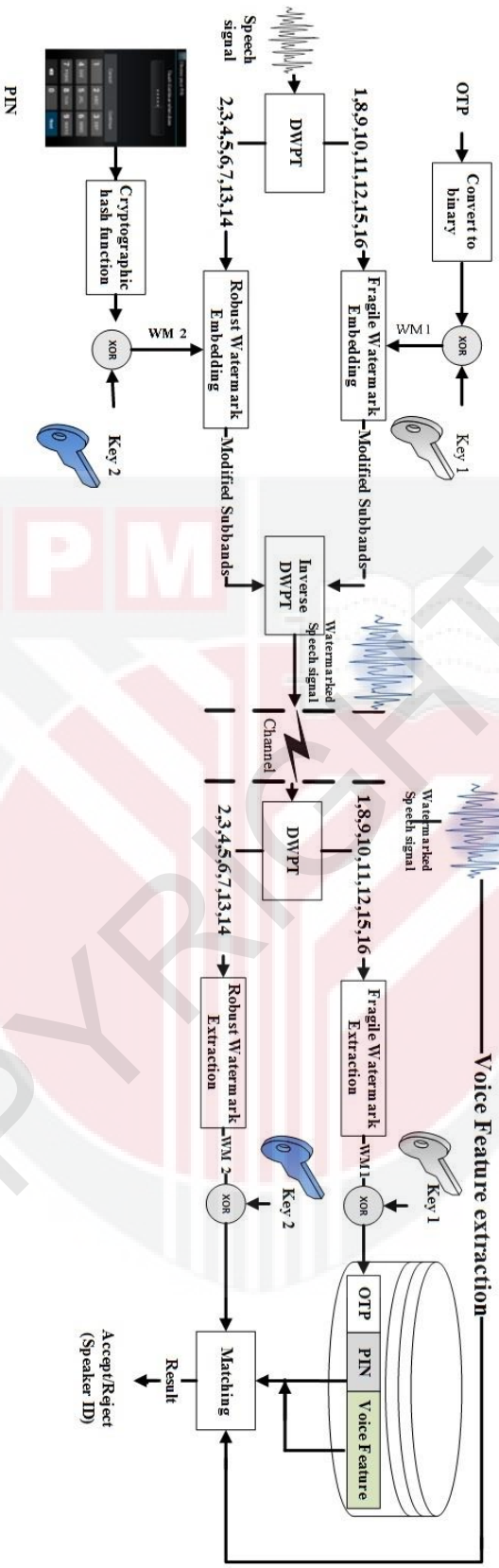


Figure 3.4 . Block diagram of proposed double digital speech watermarking for online speaker recognition system

3.2.3.1 Robust Digital Speech Watermarking Algorithm

The manipulation of digital speech can make sounds undetectable by human hearing due to advances in speech synthesizing technology. Manipulating small parts of the speech signals can also change the meaning of the whole utterances. Therefore, robust digital watermarking can be applied to speech streams in the digital world. Although different digital robust audio watermarking techniques have been proposed, available robust audio watermarking techniques cannot be applied to speech signals efficiently. The main differences between them lie in the spectral structure, temporal structure, and syntactic/semantic structure. Available robust digital speech watermarking techniques cannot satisfy the requirements such as imperceptibility, robustness, blindness, and payload at the same time due to their mutually conflicted nature and competitive nature. Furthermore, poor accuracy performance for online speaker recognition after watermarking hinders true MFA to be implemented using available robust digital speech watermarking techniques. Providing a reasonable compromise for these requirements is necessary to rectify these problems for the online MFA system based on speaker recognition systems and digital speech watermarking.

Figure 3.5 shows the critical bands which are chosen to embed the watermark. As seen in Figure 3.5, the selected bands have less speaker-specific information which has caused less degradation on the recognition performance of online speaker recognition systems. For this reason, the speech signal has decomposed into 16 critical bands by applying Discrete Wavelet Packet Transform (DWPT). Then, 8 critical bands (with numbers 2, 3, 4, 5, 6, 7, 13, and 14), where the amount of F-ratio is not much, were chosen to have minimum degradation on speaker-specific information. F-ratio curve in Figure 3.5 is captured from previous work (Lu and Dang, 2008).

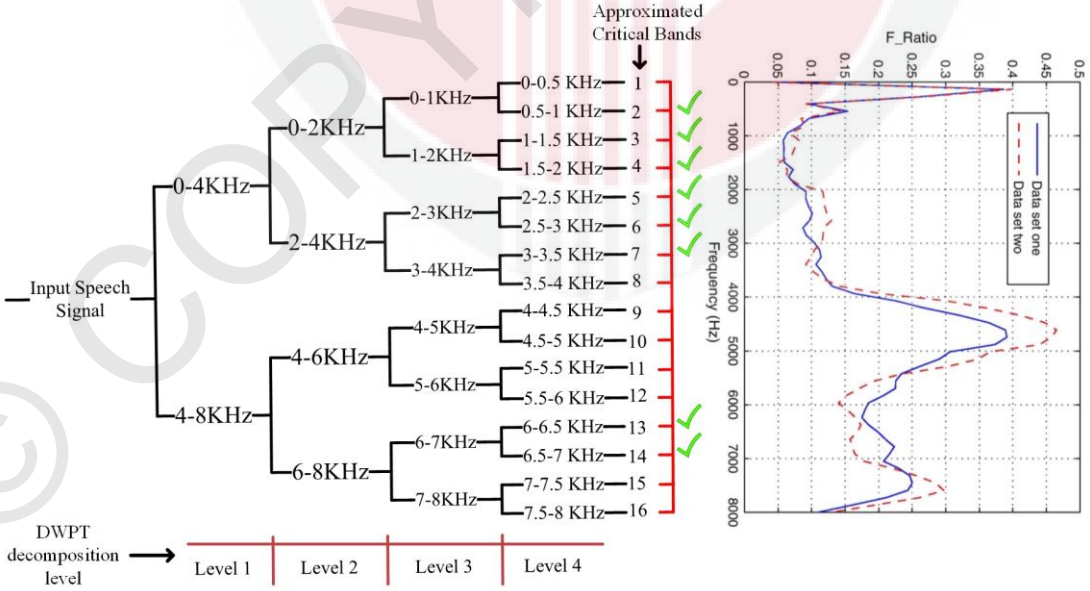


Figure 3.5. The eight less speaker-specific critical bands for watermarking by applying DWPT decomposition

In this Section a robust digital speech watermarking technique based on robust multiplicative is proposed. In this technique, the watermark is embedded by manipulating the amplitude of the speech signal (Akhaee et al., 2009). For this reason, the speech signal is segmented into none-overlapping frames with the length of *N*. Then, all the sampling of the frame is manipulated based on Equation (3.2) and Equation (3.3):

$$r_i = \alpha \times s_i \quad if \quad m_i = 1 \tag{3.2}$$

$$r_i = \frac{1}{\alpha} \times s_i \quad if \quad m_i = 0 \tag{3.3}$$

where α is the intensity of the watermark which must be slightly greater than 1, m_i is watermark bit, s_i is the original speech samples, and r_i is watermarked speech samples. Whenever α is increased, the robustness of the watermark is increased but the imperceptibility is decreased. s_i corresponds to *ith* samples of the frame. r_i is the *ith* watermarked sample of the frame.

It is demonstrated (Akhaee et al., 2009) that by knowing the watermark's strength α, variance of the noise, and variance of the original signal, it is possible to extract the watermark bit from the energy of the signal by using a predefined threshold. The detection for watermark bit is based on Equation (3.4):

$$\sum_{i=1}^{N} r_i^2 \gtrless_1^0 T \tag{3.4}$$

where *T* is the amount of threshold which is depends on the variance of the noise and signal. This detection function works well except for gaining attack. If all the samples are multiplied by a constant, the watermark bits cannot be detected at the receiver. In this thesis, a rational watermark detection technique has been applied to solve this problem. For this reason, the speech frame is divided into two sets *A* and *B* which should have equal length and energy. If their energy is not equal, then their energy can be equalized by using a distortion signal. Next, the watermark bit is embedded into *A* set based on Equation (3.2) and Equation (3.3).

For the extraction of the watermark bit from the watermarked frame, Equation (3.5) has been applied.

$$R = \frac{\sum_A r_i^{\mathrm{Order}}}{\sum_B r_i^{\mathrm{Order}}} \gtrless_0^1 T \tag{3.5}$$

Where *R* is the extracted watermark bit, *Order* is an even number and *Order*=4 is assumed to provide a tradeoff between robustness and imperceptibility.

Due to the application of DWPT, the distribution of the speech sub-bands is considered as a Generalized Gaussian Distribution (GGD) which can be assumed as Weibull distribution when DFT is applied (Akhaee et al., 2010). If GGD is assumed to be $\mu_s^2 = 0$ and σ_s^2, then it can be expressed as in Equation (3.6):

$$f_s(s; \mu, \sigma_s, v) = \frac{1}{2\Gamma\left(1+\frac{1}{v}\right)A(\sigma_s, v)} exp\left\{-\left|\frac{s-\mu}{A(\sigma_s, v)}\right|^v\right\} \tag{3.6}$$

where $\Gamma(.)$ is Gamma function which is represented by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \cong \sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x}$, v is the shape of the distribution and can be estimated based on statistical moment of the signal which is discussed briefly in Appendix A.
The amount of threshold for the detection of the watermark bit is estimated for AWGN channel. Therefore, the received watermark signal can be expressed based on Equation (3.7) and Equation (3.8).

$$r_i = \alpha \times s_i + n_i \quad if \quad m_i = 1 \tag{3.7}$$

$$r_i = \frac{1}{\alpha} \times s_i + n_i \quad if \quad m_i = 0 \tag{3.8}$$

where n_i is the noise which is added to the watermarked speech signal. Equation (3.9) estimates the probability of the watermark bit.

$$R|1 = \frac{\sum_A (\alpha \times s_i + n_i)^4}{\sum_B (s_i + n_i)^4} \Longrightarrow R|1 = \frac{\alpha^4 \sum_A s_i^4 + 4\alpha^3 \sum_A s_i^3 n_i + 6\alpha^2 \sum_A s_i^2 n_i^2 + 4\alpha \sum_A s_i n_i^3 + \sum_A n_i^4}{\sum_B s_i^4 + 4\sum_B s_i^3 n_i + 6\sum_B s_i^2 n_i^2 + 4\sum_B s_i n_i^3 + \sum_B n_i^4} \tag{3.9}$$

As seen in Equation (3.9), the amount of the detection threshold depends on the summation of the different parameters. Therefore, different series (which are considered as Normal distribution) in the nominator and denominator can be computed based on Central Limit Theorem (CLT). Although some parameters, like $\sum_A n_i^4$, are always positive and cannot be modeled by Gaussian distribution which may be negative, the probability of a negative number which is generated by this Gaussian distribution is very low due to the long length of the speech frames and big amount for μ. As a result, the mean and variance of each parameter of the nominator and denominator are estimated based on Equation (3.10) and Equation (3.10) respectively.

$$E\{\sum s_i^4\} = \sum E\{s_i^4\} = M\mu_4 \tag{3.10}$$

$$var(\sum s_i^4) = E\{(\sum (s_i^4 - M\mu_4))\}^2 = E\{(\sum (s_i^4 - \mu_4))\}^2 = \sum E\left\{\left((s_i^4 - \mu_4)\right)\right\}^2 = \sum (E\{s_i^8 - \mu_4^2\}) = M\mu_8 - M\mu_4^2 \tag{3.11}$$

where *M* is the length of each set of *A* and *B*. By assuming r=4, r=8 and based on the moment of GGD which is computed as in Appendix A, Equation (3.12) and Equation (3.13) are estimated.

$$\mu_4 = \frac{\sigma_s^4 \, \Gamma\left(\frac{1}{v}\right) \Gamma\left(\frac{5}{v}\right)}{\Gamma^2\left(\frac{3}{v}\right)} \tag{3.12}$$

$$\mu_8 = \frac{\sigma_s^8 \, \Gamma^3\left(\frac{1}{v}\right) \Gamma\left(\frac{9}{v}\right)}{\Gamma^4\left(\frac{3}{v}\right)} \tag{3.13}$$

Therefore, Equation (3.14) is estimated.

$$\sum s_i^4 \sim \mathcal{N}(M\mu_4, M\mu_8 - M\mu_4^2) \tag{3.14}$$

By assuming Gaussian signal with zero mean, Equation (3.15) can be formulated.

$$n_i \sim \mathcal{N}(0, \sigma_n^2) \implies E\{n_i^m\} = \begin{cases} 0 & for\ m = 2k+1 \\ (m-1)(m-3)\ldots \times 1 \times \sigma_n^m & for\ m = 2k \end{cases} \quad (3.15)$$

The distribution of the noise component with the moment of 4 can be estimated based on Equation (3.16).

$$\sum n_i^4 \sim \mathcal{N}(3M\sigma_n^4, 96M\sigma_n^8) \quad (3.16)$$

The rest of the components of Equation (3.9) are simply expressed as from Equation (3.17) to Equation (3.19).

$$\sum s_i^3 n_i \sim \mathcal{N}(0, M\mu_6\sigma_n^2) \quad \& \quad \mu_6 = \frac{\sigma_s^6\, \Gamma^2\left(\frac{1}{v}\right) \Gamma\left(\frac{7}{v}\right)}{\Gamma^3\left(\frac{3}{v}\right)} \quad (3.17)$$

$$\sum s_i^2 n_i^2 \sim \mathcal{N}(M\sigma_s^2\sigma_n^2, 3M\mu_4\sigma_n^4 - M\sigma_s^4\sigma_n^4) \quad (3.18)$$

$$\sum s_i n_i^3 \sim \mathcal{N}(0, 15M\sigma_s^2\sigma_n^6) \quad (3.19)$$

Therefore, by using two free auxiliary parameters *p* and *q* which are stated in Equation (3.20), R|1,p,q is expressed by Equation (3.21)

$$p = \sum_B s_i^4 \quad \& \quad q = \frac{\sum_A s_i^4}{\sum_B s_i^4} \quad (3.20)$$

$$R|1,p,q = \frac{\alpha^4 pq + 4\alpha^3 \sum_A s_i^3 n_i + 6\alpha^2 \sum_A s_i^2 n_i^2 + 4\alpha \sum_A s_i n_i^3 + \sum_A n_i^4}{p + 4\sum_B s_i^3 n_i + 6\sum_B s_i^2 n_i^2 + 4\sum_B s_i n_i^3 + \sum_B n_i^4} = \frac{u}{w} \quad (3.21)$$

where *u* and *w* are estimated based on Equation (3.22) and Equation (3.23).

$$f_U(u) \sim \mathcal{N}(\alpha^4 pq + 6\alpha^2 M\sigma_s^2\sigma_n^2 + 3M\sigma_n^4, 16\alpha^6 M\mu_6\sigma_n^2 + 36\alpha^4(3M\mu_4\sigma_n^4 - M\sigma_s^4\sigma_n^4) + 16\alpha^2 \times 15M\sigma_s^2\sigma_n^6 + 96M\sigma_n^8) \quad (3.22)$$

$$f_W(w) \sim \mathcal{N}(p + 6M\sigma_s^2\sigma_n^2 + 3M\sigma_n^4, 16M\mu_6\sigma_n^2 + 36(3M\mu_4\sigma_n^4 - M\sigma_s^4\sigma_n^4) + 16 \times 15M\sigma_s^2\sigma_n^6 + 96M\sigma_n^8) \quad (3.23)$$

For estimating the PDF of R|1,p,q , computing the density of $\frac{u}{w}$ is required. By assuming *u* and *w* as normal distribution and they are independent, Equation (3.24) can be expressed (more details in Appendix B) as:

$$f_{R|1,p,q}(r) = \int_{-\infty}^{\infty} |w| f_{U,W}(wr, w)\, dw \quad (3.24)$$

By the assumption of independent and normal distribution of *U* and *W*, $f_{U,W}(u,w)$ can be expressed as in Equation (3.25):

$$f_{U,W}(u,w) = f_U(u) \times f_W(w) \quad (3.25)$$

It should be mentioned that the closed-form solution for Equation (3.24) is available which is fully discussed in the literature and formulated as in Equation (3.26).

$$D(r) = \frac{b(r)c(r)}{a^3(r)} \frac{1}{\sqrt{2\pi}\sigma_u\sigma_w} \left[2\Phi\left(\frac{b(r)}{a(r)}\right) - 1\right] + \frac{1}{a^3(r)\pi\sigma_u\sigma_w} e^{-\frac{1}{2}\left(\frac{\mu_u^2}{\sigma_u^2} + \frac{\mu_W^2}{\sigma_W^2}\right)} \quad (3.26)$$

where each parameter is expressed as from Equation (3.27) to Equation (3.30):

$$a(r) = \sqrt{\frac{r^2}{\sigma_u^2} + \frac{1}{\sigma_w^2}} \quad (3.27)$$

$$b(r) = \frac{\mu_u}{\sigma_u^2} r + \frac{\mu_w}{\sigma_w^2} \quad (3.28)$$

$$c(r) = exp\left\{\frac{1}{2}\frac{b^2(r)}{a^2(r)} - \frac{1}{2}\left(\frac{\mu_u^2}{\sigma_u^2} + \frac{\mu_w^2}{\sigma_w^2}\right)\right\} \quad (3.29)$$

$$\Phi(r) = \int_{-\infty}^{r} \frac{1}{\sqrt{2\pi}}\, e^{-1/2u^2}\, du \quad (3.30)$$

As a result, the density of R|1 can be formulated as in Equation (3.31):

$$f_{R|1}(r|1) = \int_L^U \int_{-\infty}^{\infty} f_{R|1,p,q}(r|1,p,q)\, f_P(p)\, f_Q(q) \quad (3.31)$$

where *L* and *U* are the lowest bound and the highest bound of the energy ratio between two sets of *A* and *B* respectively. As discussed, these two sets should be selected and somehow have equal energy approximately. This situation can be stated as in Equation (3.32):

$$L < \frac{\sum_A r_i^4}{\sum_B r_i^4} < U \quad (3.32)$$

The density of parameter *P* is expressed as in Equation (3.14). However, the density of parameter q is formulated as in Equation (3.33) which is estimated from the ratio between normal and independent distribution.

$$f_Q(q) = \frac{D(q)}{\int_L^U D(q)\, dq} \quad (3.33)$$

The probability of r|0 can be computed by using the same manner. Then, the probability of detection error can be estimated based on Equation (3.34):

$$P_e = \frac{1}{2}\int_T^{\infty} f(r|0)\, dr + \frac{1}{2}\int_{-\infty}^{T} f(r|1)\, dr \quad (3.34)$$

As the main aim of the watermark detector is the minimization of the error, the threshold is calculated as in Equation (3.35):

$$\frac{\partial P_e}{\partial T} = 0 \quad \Rightarrow \quad f_r(T|0) = f_r(T|1) \quad (3.35)$$

The amount of the threshold is experimentally computed by using simulation. In the following, the robust digital speech watermarking technique has been developed based on the statistical model which is fully described in this section. The simulation parameters were assumed as follows:

a) The frame size was assumed to be 32 ms which was equal to 0.032×Fs=512 samples. A watermark bit was embedded into each frame. It is clear that whenever the size of the original speech signal is increased, the watermark capacity is increased.
b) The required level for DWPT was assumed to be 4. The watermarked sub-bands were considered as in Figure 3.4. Daubechies' wavelet function was applied for DWPT.
c) Although the watermark's intensity (α) was changed for simulation purpose, the overall assumption was $\alpha = 1.15$.
d) For channel coding, Hamming method was used with its parameters assumed to be

n=15, k=11.

e) The threshold was assumed to be 0.95. However, the proper amount for the threshold was expected to be a number near to 1 due to the equalization of energy blocks in the developed algorithm.

Robust Digital Speech Watermarking Algorithm

As discussed, the watermark bits are embedded into the specific frequency sub-bands of DWPT. Details of the embedding and extraction process are presented in the following algorithms:

Embedding process:

a) Segment the original speech signal into frame F_i with lengths of N.
b) Apply DWPT on each frame with L levels to compute the different sub-bands.
c) Select specific frequency sub-bands in the last level and arrange them into a data sequence.
d) Divide the data sequence into two sets of A and B with equal length of $N/2$ for each sets. If these two sets have different energy, their energy is equalized by using a distortion.
e) Apply a channel coding technique to improve the robustness of the watermark bits.
f) Embed the coded watermark into A set based on multiplication which is expressed in Equation (2.13) and Equation (3.3).
g) Apply inverse DWPT to reconstruct the watermarked signal.

Figure 3.6 shows the block diagram of the embedding process in the proposed robust speech watermarking technique.

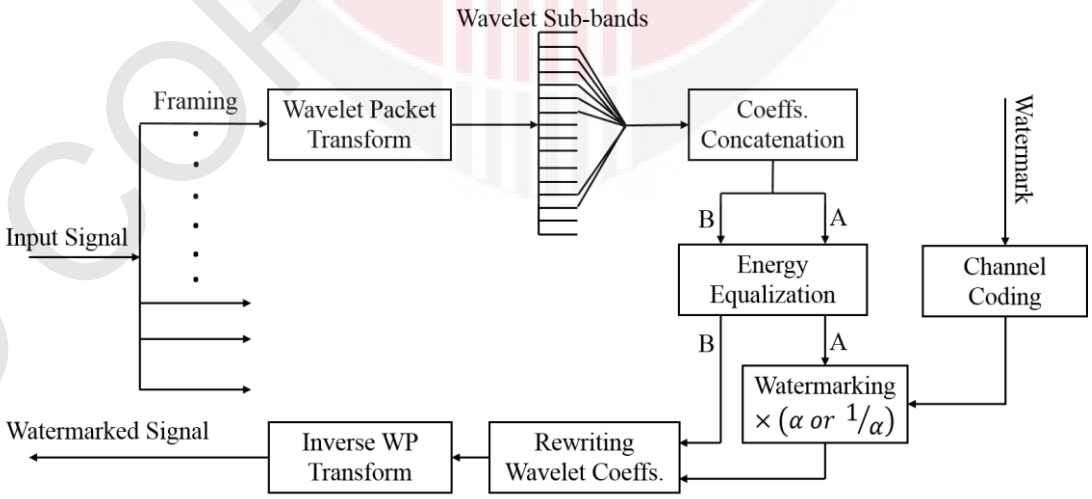


Figure 3.6. Block diagram of embedding process in the proposed robust digital speech watermarking technique in transmitter side

Extraction process:

a) Segment the watermarked speech signal into frame F_i with lengths of N (which can be considered as a public key between the transmitter and receiver).
b) Apply DWPT on each frame with L levels to compute the different sub-bands (which can be considered as a public key between the transmitter and receiver).
c) Select specific frequency sub-bands in the last level and arrange them into a data sequence.
d) Divide the data sequence into two sets of A and B with equal length of $N/2$ for each set.
e) Extract the watermark bits based on Equation (3.5).
f) Decode the watermark bits which are extracted from all embedding frames.

Figure 3.7 shows the block diagram of the extraction process in the proposed robust speech watermarking technique.

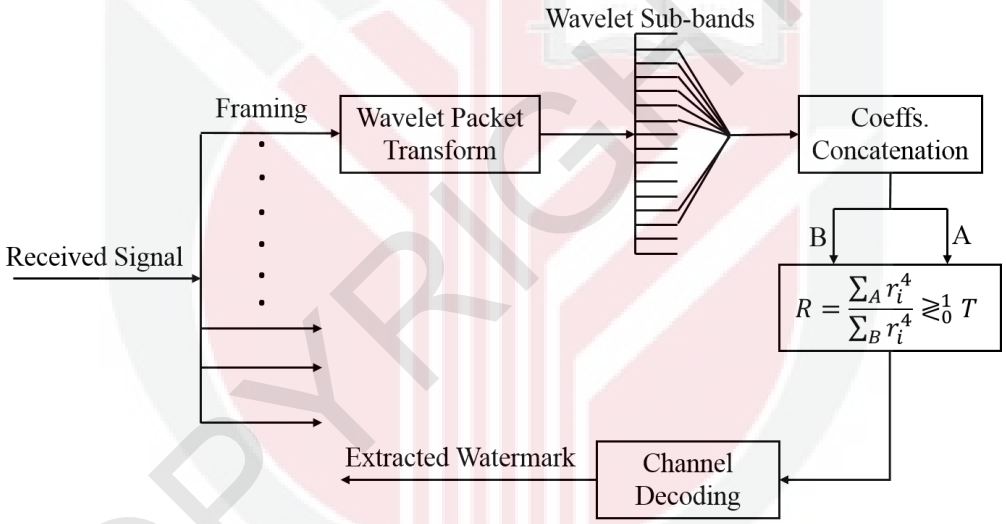


Figure 3.7.Block diagram of extraction process in the proposed robust digital speech watermarking technique receiver side

3.2.3.2 The Proposed Semi-Fragile Digital Speech Watermarking Technique

In this part, a semi-fragile speech watermarking technique has been proposed based on angle quantization (of the energy ratio between two blocks) which is very sensitive against any manipulation. This speech watermarking technique can provide authentication over an unknown channel. The proposed semi-fragile speech watermarking method can provide imperceptibility. Any manipulation of the watermark signal can also destroy the watermark bits which are changed into random bit streams. Any small manipulation of the speech signal can seriously change these angles; therefore, the quantization of the signal's angles is a good

candidate for semi-fragile speech watermarking.
In order to apply angle quantization, each watermark bit is embedded into two sets of the original signal. For this reason, two sets of the original signal (x_1 and x_2) have been selected to provide a space in two dimensional coordinate system. Then, the polar coordinate of (x_1,x_2) is calculated based on Equation (3.36) and Equation (3.37) as shown in Figure 3.8:

$$\theta = \arctan\left(\frac{x_2}{x_1}\right) \quad (3.36)$$

$$\mathrm{r} = \sqrt{x_1^2 + x_2^2} \quad (3.37)$$

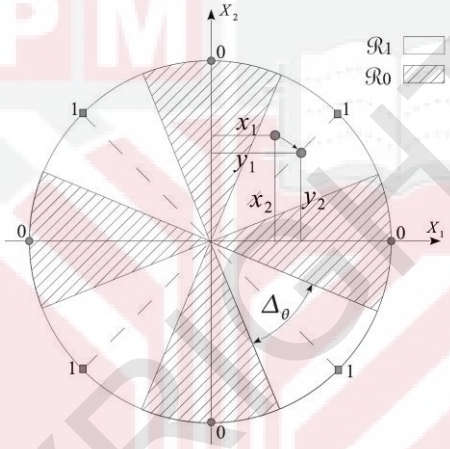


Figure 3.8. Embedding watermark by angle quantization (Coria et al., 2009)

In angle quantization, θ is quantized to embed the watermark bit. However, this technique is very fragile as even without any attack, the watermark bits cannot be extracted and hence can causes a serious error. To overcome this problem, the watermark bits are embedded by quantization of the ratio between two energy blocks of the original signal. Similar to the proposed robust digital speech watermark technique, one bit only is embedded into each frame by using semi-fragile digital speech watermark technique. However, each watermark bit is repeatedly embedded into a frame to reduce the error. Therefore, each frame is divided into blocks with lengths of L_b, and two sets of *X* and *Y* are selected. Then, θ is calculated as in Equation (3.38):

$$\theta = \arctan\left(\frac{\sum_{i=1}^{L_b/2} y_i^2}{\sum_{i=1}^{L_b/2} x_i^2}\right) \quad (3.38)$$

After angle quantization, the variation for *Y* should be estimated. In this thesis, the method of Lagrange has been applied to estimate the coefficients after angle quantization. Lagrange

method can decrease the effect of watermark distortion after angle quantization. Therefore, each watermarked coefficient is estimated by solving an optimization problem which is formulated as in Equation (3.39):

$$\begin{cases} Cost: & J(Y) = \sum_{i=1}^{L_b/2} \left(y_i^Q - y_i\right)^2 \\ Condition: & C(X) = \sum_{i=1}^{L_b/2} \left(y_i^Q\right)^2 - \theta^Q \times E_X = 0 \end{cases} \quad (3.39)$$

For solving this optimization problem, Lagrange method should estimate the optimized values of the equation system as in Equation (3.40):

$$\nabla J(Y) = \lambda \nabla C(X) \quad (3.40)$$

These optimized value are simply computed by solving the following Equation (3.41) and Equation (3.42):

$$y_i^{Q,Opt} = \frac{y_i}{1-\lambda_{Opt}} \quad (3.41)$$

$$\lambda_{Opt} = 1 - \sqrt{\frac{E_Y}{\theta^Q \times E_X}} \quad (3.42)$$

Semi-Fragile Digital Speech Watermarking Algorithm

As discussed, the watermark bits are embedded into the specific frequency sub-bands of DWPT. Details of the embedding and extraction process are presented in the following algorithms:

Embedding process:

a) Segment the original speech signal into frame *Fi* with lengths of *N*.
b) Apply DWPT on each frame with *L* levels to compute the different sub-bands.
c) Select specific frequency sub-bands in the last level and arrange them into a data sequence.
d) Divide the data sequence into different blocks with lengths of L_b. Then, each block is divided into two sets of *X* and *Y* with equal length of *N*/2 for each set.
e) Compute the energy ratio for both sets of *X* and *Y* $\frac{E_Y}{E_X}$.
f) Embed the watermark bit repeatedly into all the bocks of a frame based on Equation (3.43):

$$\theta^Q = \left\lfloor \frac{\theta + m_i \times \Delta}{2\Delta} \right\rfloor \times 2\Delta + m_i \times \Delta \quad (3.43)$$

where Δ corresponds to quantization steps, m_i is the angle of the energy ratio, and θ^Q is the modified angle of the energy ratio. Selecting small quantization steps gives more imperceptibility but less robustness and vice versa.
g) Apply Lagrange method on *Y* set to perform the required changes for minimizing the

watermarked distortion.

h) Apply inverse DWPT to reconstruct the watermarked signal.

Figure 3.9 shows the block diagram of the embedding process in the proposed semi-fragile speech watermarking technique.

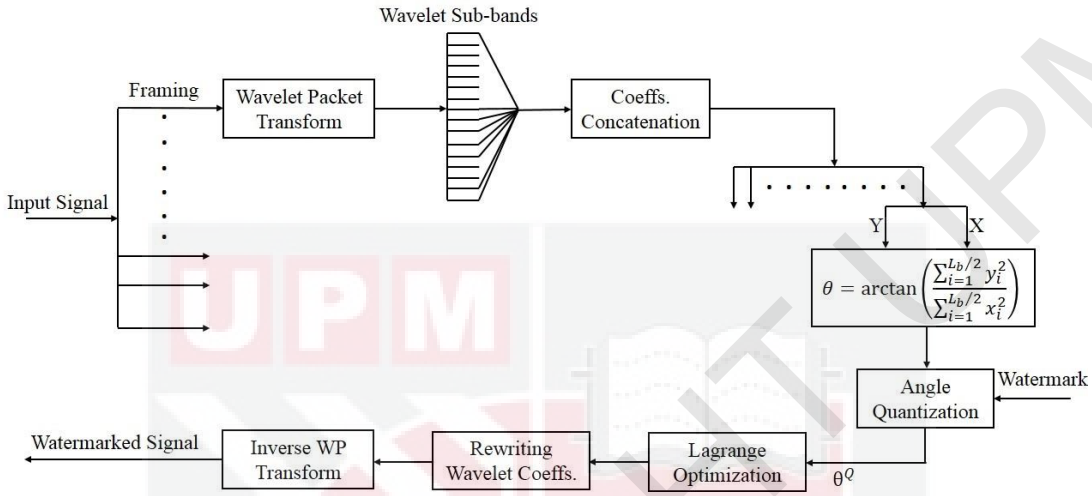


Figure 3.9. Block diagram of embedding process in the proposed fragile digital speech watermarking technique in transmitter side

By selecting a simple technique for the embedding process, the extraction process of the watermark is also reversed as described in the following**:**

Extraction process:

a) Segment the original speech signal into frame F_i with lengths of N.
b) Apply DWPT on each frame with L levels to compute the different sub-bands.
c) Select the specific frequency sub-bands in the last level and arrange them into a data sequence.
d) Divide the data sequence into different blocks with lengths of L_b. Then, each block is divided into two sets of X and Y with equal length of $N/2$ for each set.
e) Compute the energy ratio for both sets of X and Y i.e. $\frac{E_Y}{E_X}$.
f) Extract the binary watermark bit from angle θ which is the nearest quantization step to this angle based on Equation (3.44):

$$\hat{b}_k = \text{argmin}_{b_k=\{0,1\}} \left| r_k - Q_{b_k}(r_k) \right| \tag{3.44}$$

Where is r_k the angle of the energy ratio of the received signal, Q_{b_k} is the quantization function while meeting the watermark bits $b_k = \{0,1\}$.

g) Perform step e and step f repeatedly for all the blocks of a frame.
h) By embedding the same watermark bit in each block of a frame, different bits are extracted from the frame which must be made them into one bit. For this reason, a

threshold has been considered to decide regarding the extracted bit. When the number of the extracted bits for 1 is higher than the threshold, the extracted watermark bit is 1. Otherwise, the number of 0 bits is higher than 1 and the extracted watermark bit is 0. Whenever the threshold is considered to be near to 1, the fragility of the developed semi-fragile system is increased. However, when this threshold is near to 0.5, the robustness of the developed semi-fragile system is increased.

Figure 3.10 shows the block diagram of the extraction process in the proposed semi-fragile speech watermarking technique.

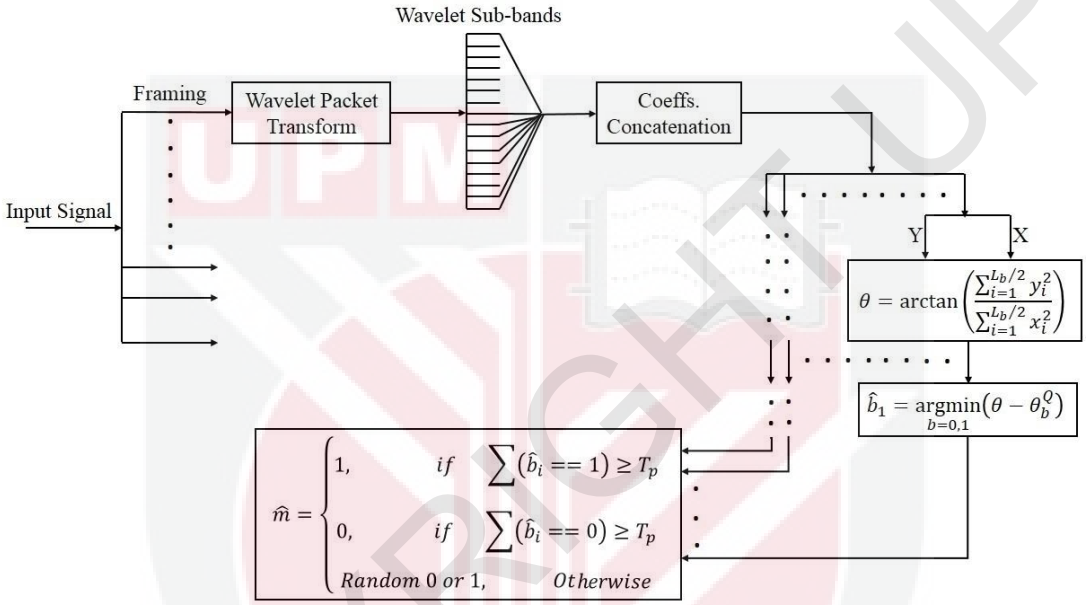


Figure 3.10. Block diagram of extraction process in the proposed semi-fragile digital speech watermarking technique in receiver side

The simulation parameters were assumed as follows:

a) The size of the frames was the same as the robust digital speech watermarking to preserve the integrity for using both watermarking techniques simultaneously. Therefore, the size of each frame was 32 ms which was equal to Fs×0.032=512 of the samples.
b) The level of the wavelet was 4. The selected sub-bands for watermarking were explained in Figure 3.4. Daubechies' wavelet function was also used for DWPT.
c) The size of each block in the frame was considered as 8 and the size of each set of *X* and *Y* in the block was equally divided into 4.
d) To preserve fragility and imperceptibility, the quantization step was assumed to be $\Delta = \pi/64$. Whenever the quantization step was increased, the fragility of the watermark was decreased. Furthermore, increasing the quantization step could decrease the imperceptibility of the speech signals in terms of SNR.
e) The decision threshold for the extraction of the watermark bits was assumed as 0.9. Whenever the threshold was increased to 1, the fragility of the developed semi-fragile system was increased. However, if this threshold was decreased to 0.5, the robustness

of the developed semi-fragile system was increased. For serious noise (SNR=0 dB), it appeared that the threshold could not affect the fragility of the watermark as the watermark bits were extracted in a random sequence.

3.3 Development of an Algorithm for Online Speaker Recognition Systems Based on Digital Speech Watermarking

As mentioned in Chapter 1, the watermark can be a proper candidate for protecting channel communication especially for digital purposes. The watermark can also be used as a header for telephone recording in the digital case, like the header in the analog case, for demonstrating that the signal has not been tampered with in the court. The speech watermark can be embedded as marks or time-stamps inside the speech signal to prevent its features from any channel modification like intentional or unintentional manipulation. Therefore, the watermark can be used for ownership identification and verification through fingerprinting, transactional, and proof of authentication of the watermark's characteristics.

As discussed, the speaker's password (or PIN) must be registered in the system's database during the enrolment step. During the testing phase, when the signal is clear from any watermark, the hash and encrypted PIN is embedded as the watermark into the speech signal which is pronounced by the speaker at the transmitter's side. When the watermarked speech is passed through the communication channel, the watermark is extracted and decrypted from the speech signal. Finally, this watermark is applied as the speaker PIN and it checks with the system's database. By using this PIN, the speaker is verified or identified. The matching between the extracted PIN from the database and the watermarked speech signal is done by the matcher based on BER. This BER can be used as the threshold for evaluation the developed system. It is noted that the proposed technique should be applied as a complimentary technique (as MFA method) with conventional speaker recognition by using the speaker frame selection technique for improving accuracy, recognition performance, and robustness.

3.3.1 Speaker Identification Process Based on Digital Watermark

For speaker identification evaluation, each speaker ID is considered as a PIN for that speaker. Each speaker ID is hashed and converted to a binary code. Then, this binary code is encrypted by XOR operation with a key and embedded into every wave of the speaker on the transmitter's side. In the receiver side, the extracted binary code is detected from the watermarked signal and it is decrypted and matched with the hash of the speaker's IDs in the speaker's database which is already registered. It is possible to select two thresholds for identification purpose. The speaker ID, with the minimum computed BER in respect to the extracted speaker's ID, is selected as the identified speaker. This minimum BER with the determined level can match the test speakers and claimer speakers.

For more security, it is possible to compare the minimum computed BER with a predefined threshold. If the minimum computed BER is lesser than the predefined threshold, it is identified. Otherwise, it is rejected. This threshold helps to separate the claimers from the impostors.

3.3.2 Speaker Verification Process Based on Digital Watermark

For speaker verification, the binary PIN is used to evaluate the performance of speaker verification system in terms of EER. The binary PIN is hashed, encrypted, and embedded into every wave of the speaker. Then, the extracted binary bits are extracted from the watermarked signal. Next, they are decrypted and compared to the result of the hashing of the original binary bits by means of BER. If the computed BER is less than the predefined threshold, the speaker is accepted and verified. Otherwise, it is rejected.

3.4 A General MFA Model Based on Online Speaker Recognition and Digital Speech Watermarking

In this part, an MFA model was developed based on online speaker recognition and digital speech watermarking technology. Three phases such as sign up, login, and recognition are discussed in detail. In addition, the possibility of changing the PIN is discussed. For better explanation, Table 3.1 presents the notations of the proposed MFA model which is shown in Figure 3.11.

Table 3.1. Applied notations for proposed MFA model.

Symbol	Notation
OSRS	Online speaker recognition system.
SPKR	Speaker.
SPKi	Voice of the speaker.
IDi	Identity of the speaker.
BMi	Voice biometric feature of the speaker.
PWi	PIN selected by the speaker.
$\Theta_1, \Theta_2, \Theta_3$	Thresholds for voice biometric and watermarking systems.
Key1	Private Key shared between SPKR and OSRS.
Key2	Private Key shared between SPKR and OSRS.
OTP	One Time Password sent by online speaker recognition system to the speaker.
$\oplus$	XOR operation.
$WM\_EX(.)$	Watermark extraction process.
$WM\_EM(.)$	Watermark embedding process.
Hash(.)	One-way hash function.
VFE(.)	Extract the voice biometric feature from the speech signal.
VFM(.)	Model the voice biometric feature for SPKRi.

VFS(.) Compute the score.

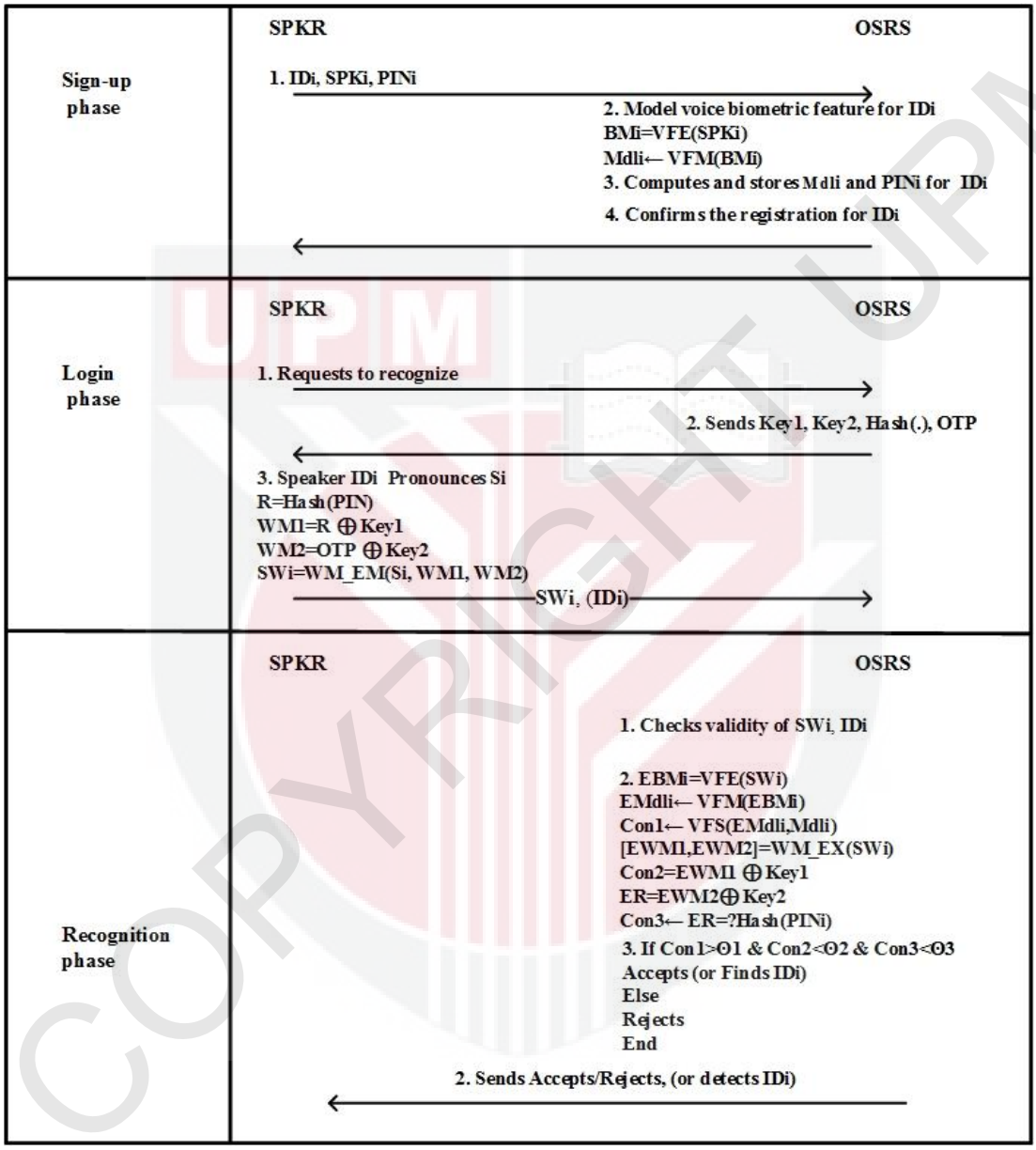


Figure 3.11 . The proposed MFA model.

Each phase of **Figure 3.11** is explained in detail in the following:

a) Sign-up phase

Before login, the speaker must register himself or herself in the system. This phase can be done in front of the system or via a secure channel. The speaker needs to do the following

steps:
Step 1: First, the speaker (SPKRi) provides his or her identity (IDi), voice (SPKi), and selects a PIN (PINi) personally.
Step 2: Then, the system (OSRS) computes BMi and PINi as follows:
BMi=VFE(SPKi)
Mdli=VFM(BMi)
Step 3: The system (OSRS) saves Mdli and PINi.
Step 4: Finally, the speaker (SPKRi) is registered to the system (OSRS) through IDi, Mdli, and PINi.

b) Login phase

When the speaker needs to be recognized by the MFA system, he or she must do the following steps:

Step 1: Firstly, the speaker (SPKRi) requests to be recognized by the system (OSRS). Then, he or she receives Key1, Key2, Hash(.) and OTP from the online system.

Step 2: Next, the speaker (SPKRi) pronounces a sentence as Si and enters his or her PIN (C_PINi) in the system (OSRS), and enter OTP.

Step 3: The speaker has to perform the following operations:

$$R = Hash(\text{C\_PINi})$$
$$WM1 = R \oplus Key1$$
$$WM2 = OTP \oplus Key2$$
$$SW\text{i} = WM\_EM(\text{Si}, WM1, WM2)$$

Step 4: Finally, the speaker (SPKRi) sends the watermarked speech signal (SWi) during the identification process. Apart from SWi, the speaker (SPKRi) should send his or her claim (IDi) in the verification process.

c) Recognition phase

While a request (SWi) is received by the system (OSRS), the following steps must be performed:
Step 1: First, the system (OSRS) checks the validity of the request (SWi) for speaker identification and the speaker (IDi) for speaker verification purpose.
Step 2: When Step1 is valid, then the following operation must be done:

$$\text{EBMi} = \text{VFE}(SW\text{i})$$
$$\text{EMdli} = \text{VFM}(\text{EBMi})$$
$$\text{Con1} \leftarrow \text{VFS}(\text{EMdli}, \text{Mdli})$$
$$[EWM1, EWM2] = WM\_EX(SW\text{i})$$
$$Con2 \leftarrow EWM1 \oplus Key1$$
$$ER = EWM2 \oplus Key2$$
$$Con3 \leftarrow ER \stackrel{?}{=} Hash(\text{C\_PINi})$$

Step 3: Check the following conditions:

$$If\ Con1 > \Theta_1\ \&\ Con2 < \Theta_2 \&\ Con3 < \Theta_3$$

Accept the speaker (SPKRi) with the speaker IDi for speaker verification.
Identify the speaker (SPKRi) with the identity of IDi for speaker identification.
Else
Reject the speaker (SPKRi) with the speaker IDi for speaker verification.
Unable to identify the speaker (SPKRi) with the identity of IDi for speaker identification.
End

d) Change PIN:

In another situation, the speaker (SPKRi) can change his or her PIN freely. For this purpose, the following steps are performed to change old PINi to new $\widehat{\text{PIN}}_i$:
Step 1: First, the speaker (SPKRi) requests to change his or her password to the system (OSRS).
Step 2: The system (OSRS) sends (Hash(.), Key1, Key2) to the speaker (SPKRi).
Step 3: The speaker (SPKRi) provides his or her identity (IDi), voice (Si), and enters the old password (PINi) personally. The PINi is secured by key1.
M1 =Hash(PINi)
PIN_old $= M1 \oplus Key1$
M2 =Hash($\widehat{\text{PIN}}_i$)
PIN_new $= M2 \oplus Key2$
Step 4: The speaker (SPKRi) sends his or her request (IDi, PIN_old, PIN_new, Si) through a secure channel.
Step 5: Next, the following operations are performed to verify the identity of the speaker (SPKRi) in the system (OSRS):
$\text{EBMi} = \text{VFE(Si)}$
$\text{EMdli} = \text{VFM(EBMi)}$
$\text{Con1} \leftarrow \text{VFS(EMli, Mdli)}$
$R1 = \text{PIN\_old} \oplus Key1$
$Con1 \leftarrow R1 \stackrel{?}{=} Hash(\text{C\_PINi})$
Step 6: Check the following condition:
$If\ Con1 > \Theta_1 \& Con2 < \Theta_2$
$\text{R2} = \text{PIN\_new} \oplus Key2$
Replace PINi with R2 in the system (OSRS).
Else
Reject the request for PIN change.
End

3.4.1 Threat Model

In order to develop the MFA model based on digital speech watermarking and online speaker recognition, the most important issue is analyzing the security of the proposed MFA model. However, the definition of security should be clarified. Therefore, the security of the proposed MFA model is discussed in two main parts. Firstly, the security requirement of the MFA model, which is the main goal to achieve, has been discussed. Secondly, the attacker model, which is defined as the potential attack that the MFA model is dealing with, has been discussed.

3.4.1.1 Security Requirements of the Proposed MFA Model

Based on the main requirements of the proposed MFA model, two applications of the digital speech watermarking are discussed in the following:

a) Fingerprinting: This application is useful to identify the legitimate speaker who pronounces the speech signal. The ownership of the speech signal should be tractable even when an adversary seriously collude with the speech signal. The adversary should also not be able to easily create an ambiguity for the legitimate speaker when detecting his or her fingerprints which have already been embedded into the speech signal. To achieve this watermark property, a robust digital speech watermarking should be applied. Then, S is the speaker, C is the speech signal, A is an adversary, $Dist(.,.)$ is the perceptual distance measurement between two speeches, $T(.)$ is the tracing function for detecting the watermark and t is the threshold. This property can be formally defined as:

   Definition: For the fingerprinting speech signal C, $C \in S$ is robust against any adversary attack $J=A(C)$ such that $Dist(C,J)<t$. When an efficient function $T(.)$ is available, then $T(J) \in S$.

b) Tamper detection: This application is useful to check the originality of the speech signal. A receiver of the speech signal should be assured that the speech signal has not be tampered with by an unauthorized party. To achieve this watermark property, a semi-fragile digital speech watermarking should be applied. When $CHL(.)$ is the unintentionally degradation function (i.e., channel effect) on the speech signal, $V(.)$ is the efficient tamper-proofing verification function which extractes the semi-fragile watermark.
   Definition: For the authenticator speech signal C , $C \in S$ is free from any tampering when both the following conditions are found:
   I. For any negligible effect on the speech signal $C'=CHL(C)$ such that $Dist(C,C')<t$, then $Prob[V(C')=Yes]<£1$.
   II. For any adversary A and the speech signal J=A(C) such that Dist(C,J)<t, then $Prob[V(J)=Yes]>£1$.

   These two statements show that when the speech signal is tampered intentionally, then the probability of the tamper detection is higher than the threshold. However, when the speech signal is just manipulated unintentionally, then the probability of the tamper detection is less than the threshold which may be negligible.

3.4.1.2 Attacker Model for the Proposed MFA Model

Apart from the security requirements for the developed MFA model, it is crucial to have a look at attacks dealt with for proposed MFA model. A suitable model attacker may properly improve the security of the system. The attacker model can detect the potential vulnerable point in the proposed MFA model. Although it cannot predict which kind of attacks have been used by an adversary, it can require rigorous treatment for the potential attacks. In the

following, two categories of the attacks have been discussed including general attacks and signal processing attacks.

(a) General Attacks

Guessing attack: It is highly desirable for an MFA system to be secure in terms of guessing attack or exhaustive search attack. Actually, a guessing attack means increasing FAR of the online speaker recognition system. This increase can be done by brute force search of an adversary which may record or synthesize the speech. By using the False Match Rate (FMR) for the result (O'Gorman, 2003), the keyspace for the speech is between $\frac{1}{0.007} = 142.9$ and $\frac{1}{0.0003} = 3333.3$. Furthermore, a 20-bits PIN has the keyspace of $2^{20} = 1048576$.
It can be seen that PIN(1048576)>Speech(3333.3). Although none of the keyspace of the PIN and speech is large enough to be secure against guessing attack and exhaustive search attack, employing both PIN and speech biometric can be sufficient. As a result, the guessing combination of PIN and speech cannot be easy. A large keyspace can defend the MFA system against these attacks. Apart from this situation, the adversary can extract the watermark from the speech signal. Even when he or she can extract the watermark, it is just a secure message as a result of hashing PIN with the encrypted speech by a key $(hash(PIN) \oplus key)$. Therefore, the adversary needs to have both key and hashing function.

Plain text or template attack: Plain text or template attack mainly happen on the online speaker recognition side. An adversary can attack the speech biometric when the speech is not a secret. Therefore, speech template protection is not fully achieved for online speaker recognition. The best way to assure this security is to authenticate the speech that is captured in a lively way which is not already entered as a file. An OTP can improve the security of the MFA system which reveals any manipulation of the speech template.

Eavesdropping, Theft, and Copying Attacks: One of the threats is to steal the PIN. This threat may be done by eavesdropping attack which requires an adversary to have a physical presence. Using the combination of speech biometric, PIN, and OTP as an MFA system is a good defense against this attack because the adversary needs to steal all of these factors. Furthermore, the theft and copying attacks are difficult because OTP and PIN are hidden in the speech signal. In addition to watermarking technology to protect the speech biometric template from theft and copying attacks, using Exclusive OR (*XOR*) operation as an encryption can secure the MFA system.

Counterfeiting or spoofing attack: Similar to theft and copy attacks, forgery attack can threaten the speech biometric at the sensor part. Although the speech biometric can be replaced easily and it does not have secrecy, the communication channel security of the online speaker recognition system can be protected by a combination of robust and semi-fragile speech watermarking.

Replay attack: In a replay attack, an adversary tries to insert the speech signal on the channel between the speaker and speaker recognition system. Even when the speech

is encrypted, the adversary can still put the encrypted data on the channel. Furthermore, when speech is sent directly, the speech signal can be replayed. The main defense mechanism is verification of the legitimacy of the speech signal which is successfully done by digital speech watermarking. Robust speech watermarking can ensure that the adversary does not alter the speech signal. Furthermore, using OTP as time stamps can resist against a replay attack. At any time a delay for transmitting the watermarked speech signal can be terminated by the MFA system.

Trojan horse attack: The Trojan horse attack tries to masquerade as a trust application to gain the information of the speaker. This attack is used to steal PIN and speech biometric. The main defense is assurance about the legitimacy and trust of the authenticator capture sensor. There is not much effort can that be done when the speaker wants to have his or her speech biometric and PIN in a Trojan horse. However, using OTP can help the MFA system not to succumb to this attack. When the speech biometric is replaced by a speech signal containing the Trojan horse to produce yes-match for anyone, the adversary cannot produce PIN and OTP.

Denial-of-service attack: In some conditions, an adversary tries to increase FRR to force the system to lockout to limit the number of the adversary's attempts. Such a service attack can be defended by combining the speech biometric and PIN as an MFA system. In the MFA system, it is not possible for the adversary to simply make any number of incorrect attempts.

Session hijack: In some situations, the previously valid watermarked speech signal may be recorded to exploit for an unauthorized access which is known as session hijacking. For every login session, a unique OTP is embedded as a timestamp into the speech signal. The uniqueness of each session can guarantee the freshness of the property (Chaturvedi et al., 2013).

Man-in-the-middle: The recorded speech at the sensor should pass through many online speaker recognition components. Therefore, the reliability of the system's integrity is necessary against man-in-the-middle attack (Roberts, 2007). Using two keys and hash function can protect the watermark from any misuse. Furthermore, the watermark is tied intrinsically to the speaker biometric which can prevent the adversary from injecting the compromised PIN. Any adversary's attempt can be detected by using the semi-fragile watermark.

Non-repudiation: In some conditions, there is a requirement that the sender cannot deny sending the speech signal to the receiver. Sometimes the sender may have the ability to deny which is known as plausible deniability (Li et al., 2006). This type of attack can be protected by a combination of the speech biometric, OTP, and PIN as an MFA system. As a result, it is difficult for the speaker to deny because three factors such as his or her speech biometric, PIN, and timestamp as OTP are available in his or her speech signal at the same time.

(b) Signal Processing Attacks

Sometime an adversary tries to apply a signal processing operation to remove the watermark's signature in the speech signal. These attacks consist of adding noise, filtering, and compression. They also perform some distortions to the speech signal. It is important to provide security against signal processing attacks since the formulation of such security is difficult. In addition, designing the digital speech watermarking which can resist against all possible signal processing operations is very hard.

3.4.2 Attack Analysis of the Proposed MFA Model

The amount of risk and threat for online single factor authentication methods increases due to a lack of security in ordinary ID and password systems. These systems are vulnerable against malware attacks, replay attacks, offline brute force attacks, key logger Trojans, dictionary attacks, and shoulder surfing. Recently, 2 Factor Authentication (2FA) has become a mandatory demand in many governmental policies (Kim and Hong, 2011). Four levels of assurance are defined by the Office of Management and Budget (OMB 04-04) as in (Bolten, 2003). Each level shows the degree of confidence that the user is in fact a legitimate user. Table 3.2 presents these four levels:

Table 3.2.Levels of assurance in authentication systems based on OMB 04-04.

Level	Secret	Definition
Level 1	Any type of token with no identity proof	Low confidence assurance available in identifier technique.
Level 2	Single factor authentication with some identity proof	Medium confidence assurance available in identifier technique.
Level 3	MFA with stringent identity proof	High confidence assurance available in identifier technique.
Level 4	MFA+ crypto token with registration per person	Very high confidence assurance available in identifier technique.

As seen in Table 3.2, applying cryptographic hash function, speaker voice biometric, and digital speech watermarking has improved the developed MFA system in level 4. Furthermore, the registration of speaker specific features and PIN in the enrolment for each user is capable of the developing the MFA system to have enough protection against different attacks.

Due to the diversity in proposing user the authentication method, a standard is defined by presenting five levels of user authentication (Kim and Hong, 2011) Table 3.3 shows these five level:

Table 3.3.Five levels of user authentication (Kim and Hong, 2011; Kumar and Lee, 2013).

Level	Description
Level 1	Uses offline registration of identification information such as PIN, OTP, etc.
Level 2	Uses a soft token which is issued based on a reliable identification of the user. This reliable identification has already been done by the government with by passport, driver's license, etc.
Level 3	Uses combination of an accredited certificate (a soft token) with other security factors such as mobile phone, security card, security token, etc.
Level 4	Uses combination of an accredited certificate (a soft token) with other hardware security devices like OTP.
Level 5	Uses combination of an accredited certificate (a soft token) with watermarked biometric information like key with fingerprints.

As discussed in Section 3.4.1.2 and Table 3.3, it can be concluded that the proposed MFA system has protection against various attacks, as summarized in Table 3.4. Therefore, the proposed MFA system is in level 5.

Table 3.4. Required authentication protection mechanism for each level (Kim and Hong, 2011; Kumar and Lee, 2013).

Required Protection	Level 1	Level 2	Level 3	Level 4	Level 5
Online guessing	Yes	Yes	Yes	Yes	Yes
Replay	Yes	Yes	Yes	Yes	Yes
Eavesdropper	No	Yes	Yes	Yes	Yes
Verifier impersonation	No	No	Yes	Yes	Yes
Man-in-the-middle	No	No	Yes	Yes	Yes
Session Hijacking	No	No	No	Yes	Yes
Signer impersonation	No	No	No	No	Yes

3.5 Speaker Recognition Test Bed

In this experiment, a series of MATLAB Toolbox were used to provide a test bed for speaker verification and speaker identification. For this reason, MSR Identity MATLAB Toolbox v1.0 (Seyed Omid Sadjadi, 2013) was used to construct speaker verification system. Voice Box MATLAB Toolbox (Brookes) was applied to the speaker identification system. For hash function, DataHash MATLAB function (Simon, 2012) was applied. For the performance evaluation of the speaker verification system, two state-of-the-art paradigms were used, including GMM-UBM (Reynolds et al., 2000) and i-vector (Dehak et al., 2011; Kenny, 2012) based speaker verification systems. For the performance evaluation of speaker identification, only GMM speaker identification system (Pathak and Raj, 2013; Reynolds, 1995) was constructed. Other systems such as digital speech watermarking and speaker frame selection were implemented specifically as MATLAB codes.

3.6 Summery

This chapter has been discussed the MFA framework of this thesis which is based on speech watermarking and speaker recognition systems. To achieve this aim, various parts have been developed such as frame selection, robust speech watermarking, fragile speech watermarking, and MFA model. As discussed, the developed framework is in level 5 due to use registration per person, cryptography, MFA model. Furthermore, threat model and attack analysis have been discussed.

REFERENCES

Abdulla, W. H. (2014). *Audio Watermark: A Comprehensive Foundation Using Matlab*: Springer.

Akhaee, M. A., Kalantari, N. K., Ahadi, S. M., & Amindavar, H. (2009). Robust multiplicative patchwork method for audio watermarking. *Audio, Speech, and Language Processing, IEEE Transactions on, 17(6), 1133-1141.*

Akhaee, M. A., Kalantari, N. K., & Marvasti, F. (2009). *Robust multiplicative audio and speech watermarking using statistical modeling.* Paper presented at the IEEE International Conference on Communications, 2009. ICC'09. , 1-5.

Akhaee, M. A., Kalantari, N. K., & Marvasti, F. (2010). Robust audio and speech watermarking using Gaussian and Laplacian modeling. *Signal Processing, 90*(8), 2487-2497.

Akhaee, M. A., Amini, A., Ghorbani, G., & Marvasti, F. (2010, March). *A solution to gain attack onwatermarking systems: logarithmic homogeneous rational dither modulation.* IEEE International Conference on In Acoustics Speech and Signal Processing (ICASSP), 2010 (pp. 1746-1749). IEEE.

Al-Haj, A. (2014). An imperceptible and robust audio watermarking algorithm. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(1), 1-12.

Al-Haj, A., Mohammad, A., & Bata, L. (2011). DWT-Based Audio Watermarking. *International Arab Journal of Information Technology (IAJIT), 8*(3), 326-333.

Al-Haj, A., Twal, C., & Mohammad, A. (2010). *Hybrid DWT-SVD audio watermarking.* Paper presented at the, 2010 Fifth International Conference on Digital Information Management (ICDIM), 525-529.

Al-Nuaimy, W., El-Bendary, M. A., Shafik, A., Shawki, F., Abou-El-azm, A. E., El-Fishawy, N. A., . . . Abd El-Samie, F. E. (2011). An SVD audio watermarking approach using chaotic encrypted images. *Digital Signal Processing, 21*(6), 764-779.

Al-Shoshan, A. I. (2006). Speech and music classification and separation: a review. *Journal of King Saud University, 19*(1), 95-133.

Al-Yaman, M. S., Al-Taee, M. A., & Alshammas, H. A. (2012). *Audio-watermarking based ownership verification system using enhanced DWT-SVD technique.* Paper presented at the 9th International Multi-Conference on Systems, Signals and Devices (SSD) (pp. 1-5), 1-5.

Alegre, F., Amehraye, A., & Evans, N. (2013). *A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns.* Paper presented at the Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on, 1-8.

Alegre, F., Vipperla, R., Amehraye, A., & Evans, N. (2013). *A new speaker verification spoofing countermeasure based on local binary patterns.* Paper presented at the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon: France (2013), 5p.

Alegre, F., Vipperla, R., & Evans, N. (2012). *Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals.* Paper presented at the INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association.

Arora, S., & Emmanuel, S. (2003). Adaptive Spread Spectrum based Watermarking of Speech (pp. Poster 15): 9th National Undergraduate Research Opportunities Programme Congress.

Auckenthaler, R., Parris, E. S., & Carey, M. J. (1999). *Improving a GMM speaker verification system by phonetic weighting.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. , 313-316.

Baroughi, A. F., & Craver, S. (2014). *Additive attacks on speaker recognition.* Paper presented at the IS&T/SPIE Electronic Imaging, 90280Q-90280Q.

Bartlow, N., Kalka, N., Cukic, B., & Ross, A. (2007). *Protecting iris images through asymmetric digital watermarking.* Paper presented at the Automatic Identification Advanced Technologies, 2007 IEEE Workshop on, 192-197.

Bassia, P., Pitas, I., & Nikolaidis, N. (2001). Robust audio watermarking in the time domain. *Multimedia, IEEE Transactions on, 3*(2), 232-241.

Besacier, L., Bonastre, J.-F., & Fredouille, C. (2000). Localization and selection of speaker-specific information with statistical modeling. *Speech communication, 31*(2), 89-106.

Bhat, V., Sengupta, I., & Das, A. (2010). An adaptive audio watermarking based on the singular value decomposition in the wavelet domain. *Digital Signal Processing, 20*(6), 1547-1558.

Bhat, V., Sengupta, I., & Das, A. (2011). An audio watermarking scheme using singular value decomposition and dither-modulation quantization. *Multimedia Tools and Applications, 52*(2-3), 369-383.

Bhat, V., Sengupta, I., & Das, A. (2011). A new audio watermarking scheme based on singular value decomposition and quantization. *Circuits, Systems, and Signal Processing, 30*(5), 915-927.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S.Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing, 2004*, 430-451.

Blackledge, J., & Farooq, O. (2008). Audio data verification and authentication using frequency modulation based watermarking. *International Society for Advanced Science and Technology, Journal of Electronics and Signal Processing, 3*(2), 51-63.

Blamey, P., Dowell, R., Clark, G. M., & Seligman, P. (1987). Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant. *The Journal of the Acoustical Society of America, 82*(1), 38-47.

Bolten, J. B. (2003). E-authentication guidance for federal agencies. *Office of Management and Budget,(December 16, 2003). http://www. whitehouse. gov/omb/memoranda/fy04/m04-04. pdf*.

Brookes, M. VOICEBOX: A speech processing toolbox for MATLAB. 2006.

Chaturvedi, A., Mishra, D., & Mukhopadhyay, S. (2013). Improved Biometric-Based Three-factor Remote User Authentication Scheme with Key Agreement Using Smart Card *Information Systems Security* (pp. 63-77): Springer.

Chen, S., & Leung, H. (2006). *Concurrent data transmission through PSTN by CDMA.* Paper presented at the Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on, 4 pp.-3004.

Cheng, Q., & Sorensen, J. (2001). *Spread spectrum signaling for speech watermarking.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 1337-1340.

Chu, W. C. (2004). *Speech coding algorithms: foundation and evolution of standardized coders*: John Wiley & Sons.

CISCO. (2007). Wideband Audio and IP Telephony Experience Higher-Quality Media. *white paper, USA*.

Coria, L., Nasiopoulos, P., & Ward, R. (2009). A region-specific QIM-based watermarking scheme for digital images. Paper presented at the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2009. BMSB'09. , 1-6.

Costa, M. H. (1983). Writing on dirty paper (corresp.). *Information Theory, IEEE Transactions on, 29*(3), 439-441.

Cox, I. J., Doërr, G., & Furon, T. (2006). Watermarking is not cryptography *Digital Watermarking* (pp. 1-15): Springer.

Cox, I. J., Miller, M. L., Bloom, J. A., & Honsinger, C. (2002). *Digital watermarking* (Vol. 53): Springer.

Cvejic, N., Keskinarkaus, A., & Seppanen, T. (2001). *Audio watermarking using m-sequences and temporal masking.* Paper presented at the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, 2001 227-230.

De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I. (2012). Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Trans. Audio Speech Language Process, 20*(8), 2280–2290.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). *Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.* Paper presented at the INTERSPEECH, 1559-1562.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on, 19*(4), 788-798.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society, 39*(1), 1-38.

Deng, Z., Yang, Z., Shao, X., Xu, N., Wu, C., & Guo, H. (2007). Design and implementation of steganographic speech telephone. *Advances in Multimedia Information Processing–PCM 2007*, 429-432.

Drugman, T. (2011). *Advances in glottal analysis and its applications.* PhD Thesis, University of Mons, Belgium.

Drugman, T. (2012). GLOttal Analysis Toolbox (GLOAT). *(Downloaded November 2014)*.

El-Samie, F. E. A. (2011). *Information Security for Automatic Speaker Identification*: Springer.

Fatindez-Zanuy, M. (2004). On the vulnerability of biometric security systems. *IEEE Aerospace and Electronic Systems Magazine (June 2004)*, 3-8.

Faundez-Zanuy, M. (2010). Digital watermarking: new speech and image applications. *Advances in Nonlinear Speech Processing*, 84-89.

Faundez-Zanuy, M., Hagmüller, M., & Kubin, G. (2006). Speaker verification security improvement by means of speech watermarking. *Speech communication, 48*(12), 1608-1619.

Faundez-Zanuy, M., Hagmüller, M., & Kubin, G. (2007). Speaker identification security improvement by means of speech watermarking. *Pattern Recognition, 40*(11), 3027-3034.

Faúndez-Zanuy, M., & Rodríguez-Porcheron, D. (1998). *Speaker recognition using residual signal of linear and nonlinear prediction models.* Paper presented at the ICSLP, 121-124.

Faundez-Zanuy, M., Lucena-Molina, J. J., & Hagmüller, M. (2010). Speech Watermarking: An Approach for the Forensic Analysis of Digital Telephonic Recordings*. *Journal of forensic sciences, 55*(4), 1080-1087.

Fazel, A., & Chakrabartty, S. (2011). An overview of statistical pattern recognition techniques for speaker verification. *Circuits and Systems Magazine, IEEE, 11*(2), 62-81.

Feustel, T. C., Logan, R. J., & Velius, G. A. (2005). Human and machine performance on speaker identity verification. *The Journal of the Acoustical Society of America, 83*(S1), S55-S55.

Flanagan, J. L. (1972). *Speech analysis: Synthesis and perception*: Springer.

Garcia-Hernandez, J. J., Nakano-Miyatake, M., & Perez-Meana, H. (2008). Data hiding in audio signal using Rational Dither Modulation. *IEICE Electronics Express, 5*(7), 217-222.

Garofolo, J. S., & Consortium, L. D. (1993). *TIMIT: acoustic-phonetic continuous speech corpus*: Linguistic Data Consortium.

Geiser, B., & Vary, P. (2008). *High rate data hiding in ACELP speech codecs.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 4005-4008.

Gopalakrishnan, K., Memon, N., & Vora, P. L. (2001). Protocols for watermark verification. *IEEE Multimedia, 8*(4), 66-70.

Gurijala, A. (2007). *Speech watermarking through parametric modeling*: ProQuest.

Hagmüller, M., Hering, H., Kröpfl, A., & Kubin, G. (2004). Speech watermarking for air traffic control. *Watermark, 8*(9), 10.

Hämmerle-Uhl, J., Raab, K., & Uhl, A. (2011). *Watermarking as a means to enhance biometric systems: A critical survey.* Paper presented at the Information Hiding, 238-254.

Hanilci, C., & Ertas, F. (2011). *Impact of voice excitation features on speaker verification.* Paper presented at the 7th International Conference on Electrical and Electronics Engineering (ELECO). II-157-II-160.

Harjito, B., Han, S., Potdar, V., Chang, E., & Xie, M. (2010). *Secure communication in wireless multimedia sensor networks using watermarking.* Paper presented at the 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST). 640-645.

Hatada, M., Sakai, T., Komatsu, N., & Yamazaki, Y. (2002). *Digital watermarking based on process of speech production.* Paper presented at the ITCom 2002: The Convergence of Information Technologies and Communications, 258-267.

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., Leino, T., & Laukkanen, A.-M. (2013). *I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry.* Paper presented at the INTERSPEECH, 930-934.

Hofbauer, K. (2009). *Speech watermarking and air traffic control.* Ph. D. dissertation, Graz University of Technology, Graz, Austria.

Hofbauer, K., Hering, H., & Kubin, G. (2005). *Speech watermarking for the VHF radio channel.* Paper presented at the proceedings of the 4th Eurocontrol innovative research workshop.

Hofbauer, K., Kubin, G., & Kleijn, W. B. (2009). Speech watermarking for analog flat-fading bandpass channels. *Audio, Speech, and Language Processing, IEEE Transactions on, 17*(8), 1624-1637.

Hu, H.-T., Chou, H.-H., Yu, C., & Hsu, L.-Y. (2014). Incorporation of perceptually adaptive QIM with singular value decomposition for blind audio watermarking. *EURASIP Journal on Advances in Signal Processing, 2014*(1), 1-12.

Hu, H.-T., Hsu, L.-Y., & Chou, H.-H. (2014). Perceptual-based DWPT-DCT framework for selective blind audio watermarking. *Signal Processing, 105*, 316–327.

Huang, H.-C., Chu, S.-C., Pan, J.-S., Huang, C.-Y., & Liao, B.-Y. (2011). Tabu search based multi-watermarks embedding algorithm with multiple description coding. *Information Sciences, 181*(16), 3379-3396.

Huang, H.-C., & Fang, W.-C. (2010). Metadata-based image watermarking for copyright protection. *Simulation Modelling Practice and Theory, 18*(4), 436-445.

Huber, R., Stögner, H., & Uhl, A. (2011). *Two-factor biometric recognition with integrated tamper-protection watermarking.* Paper presented at the Communications and Multimedia Security, 72-84.

Hyon, S. (2012). *An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution.* Paper presented at the Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012. , Asia-Pacific, 1-4.

Inamdar, V. S., & Rege, P. P. (2014). Dual watermarking technique with multiple biometric watermarks. *Sadhana, 39*(1), 3-26.

Jain, A. K., Ross, A., & Pankanti, S. (2006). Biometrics: a tool for information security. *Information Forensics and Security, IEEE Transactions on, 1*(2), 125-143.

Jain, A. K., & Uludag, U. (2003). Hiding biometric data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 25*(11), 1494-1498.

Kajarekar, S. S., & Hermansky, H. (2001). *Speaker verification based on broad phonetic categories.* Paper presented at the 2001: A Speaker Odyssey-The Speaker Recognition Workshop, 201–206.

Kenny, P. (2012). *A small foot-print i-vector extractor.* Paper presented at the Proc. Odyssey, 1-25.

Keromytis, A. D. (2010). Voice-over-ip security: Research and practice. *Security & Privacy, IEEE, 8*(2), 76-78.

Khitrov, M. (2013). Talking passwords: voice biometrics for data access and security. *Biometric Technology Today, 2013*(2), 9-11.

Kim, D.-S. (2003). Perceptual phase quantization of speech. *Speech and Audio Processing, IEEE Transactions on, 11*(4), 355-364.

Kim, J.-J., & Hong, S.-P. (2011). A Method of Risk Assessment for Multi-Factor Authentication. *Journal of Information Processing Systems (JIPS), 7*(1), 187-198.

Kinnunen, T., Karpov, E., & Franti, P. (2006). Real-time speaker identification and verification. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(1), 277-288.

Kinnunen, T., Wu, Z.-Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). *Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 4401-4404.

Kons, Z., & Aronowitz, H. (2013). *Voice transformation-based spoofing of text-dependent speaker verification systems.* Paper presented at the INTERSPEECH, 945-949.

Kubin, G., Atal, B., & Kleijn, W. (1993). *Performance of noise excitation for unvoiced speech.* Paper presented at the Speech Coding for Telecommunications, 1993. Proceedings., IEEE Workshop on, 35-36.

Kumar, A., & Lee, H. J. (2013). Multi-Factor Authentication Process Using More than One Token with Watermark Security *Future Information Communication Technology and Applications* (pp. 579-587): Springer.

Lacy, J., Quackenbush, S. R., Reibman, A. R., Shur, D., & Snyder, J. H. (1998). *On combining watermarking with perceptual coding.* Paper presented at the Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998. , 3725-3728.

Lei, B., Yann Soon, I., Zhou, F., Li, Z., & Lei, H. (2012). A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition. *Signal Processing, 92*(9), 1985-2001.

Lei, B. Y., Soon, I. Y., & Li, Z. (2011). Blind and robust audio watermarking scheme based on SVD–DCT. *Signal Processing, 91*(8), 1973-1984.

Li, D., & Kim, J. (2014). Secure Audio Forensic Marking Alogrithm Using 2D Barcode in DWT-DFRNT Domain. *International Journal of Distributed Sensor Networks, 2014*, 1-12.

Li, K.-P., & Porter, J. E. (1988). *Normalizations and selection of speech segments for speaker recognition scoring.* Paper presented at the Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on, 595-598.

Li, Q., Memon, N., & Sencar, H. T. (2006). *Security issues in watermarking applications-A deeper look.* Paper presented at the Proceedings of the 4th ACM international workshop on Contents protection and security, 23-28.

Lien, N. T. H. (2009). *Echo Hiding using Exponential Time-spread Echo Kernel and Its Applications to Audio Digital Watermarking and Speaker Recognition*. Tokyo Institute of Technology.

Liu, C.-H., & Chen, O. T.-C. (2004). *Fragile speech watermarking scheme with recovering speech contents.* Paper presented at the Circuits and Systems, 2004. MWSCAS'04. The 2004 47th Midwest Symposium on, II-165-II-168 vol. 162.

Lu, X., & Dang, J. (2008). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech communication, 50*(4), 312-322.

Ma, L., Wu, Z.-j., Hu, Y., & Yang, W. (2007). An information-hiding model for secure communication. *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, 1305-1314.

Malik, H. M., Ansari, R., & Khokhar, A. A. (2007). Robust data hiding in audio using allpass filters. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(4), 1296-1304.

Mallat, S. (2008). A wavelet tour of signal processing: the sparse way: Academic press.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 11*(7), 674-693.

Malvar, H. S. (1992). Extended lapped transforms: Properties, applications, and fast algorithms. *IEEE Transactions on Signal Processing, 40*(11), 2703-2714.

Malvar, H. S. (1992). *Signal processing with lapped transforms*: Artech House.

Mat Kiah, M., Zaidan, B., Zaidan, A., Mohammed Ahmed, A., & Al-bakri, S. H. (2011). A review of audio based steganography and digital watermarking. *International Journal of Physical Sciences, 6*(16), 3837-3850.

MathWorks, I., Swami, A., Mendel, J. M., & Nikias, C. L. (1998). *Higher-order Spectral Analysis Toolbox: for Use with MATLAB: User's Guide*: Mathworks, Incorporated.

McCool, C., Marcel, S., Hadid, A., Pietikainen, M., Matejka, P., Cernocky, J., . . . Levy, C. (2012). *Bi-modal person recognition on a mobile phone: using mobile phone data.* Paper presented at the Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on, 635-640.

McLaughlin, J., Reynolds, D. A., & Gleason, T. P. (1999). *A study of computation speed-UPS of the GMM-UBM speaker recognition system.* Paper presented at the EUROSPEECH, 1215-1218.

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE, 79*(3), 278-305.

Mishra, J., V Patil, M., & S Chitode, J. (2013). An Effective Audio Watermarking using DWT-SVD. *International Journal of Computer Applications, 70*(8), 6-11.

Mosayyebpour, S., Sheikhzadeh, H., Gulliver, T. A., & Esmaeili, M. (2012). Single-microphone LP residual skewness-based inverse filtering of the room impulse response. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(5), 1617-1632.

Motwani, R. C. (2010). *A Voice-Based Biometric Watermarking Scheme For Digital Rights Management of 3D Mesh Models.* (PhD), University of Nevada, Reno.

Narimannejad, M., & Ahadi, S. M. (2011). *Watermarking of speech signal through phase quantization of sinusoidal model.* Paper presented at the Electrical Engineering (ICEE), 2011 19th Iranian Conference on, 1-4.

Noore, A., Singh, R., Vatsa, M., & Houck, M. M. (2007). Enhancing security of fingerprints through contextual biometric watermarking. *Forensic Science International, 169*(2), 188-194.

Nosratighods, M., Ambikairajah, E., Epps, J., & Carey, M. J. (2010). A segment selection technique for speaker verification. *Speech communication, 52*(9), 753-761.

O'Gorman, L. (2003). Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE, 91*(12), 2021-2040.

Özer, H., Sankur, B., & Memon, N. (2005). *An SVD-based audio watermarking technique.* Paper presented at the Proceedings of the 7th workshop on Multimedia and security, 51-56.

Patel, R., & Shrawankar, U. (2013). Security Issues In Speech Watermarking for Information Transmission. *arXiv preprint arXiv:1304.6872*, 830-839.

Pathak, M. A., & Raj, B. (2013). Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models. *Audio, Speech, and Language Processing, IEEE Transactions on, 21*(2), 397-406.

Pati, D., & Prasanna, S. (2010). Speaker Recognition from Excitation Source Perspective. *IETE Technical Review, 27*(2).

Persky, D., & Niem, J. (2007). VoIP Security Vulnerabilities. *white paper, SANS Institute*.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition* (Vol. 14): PTR Prentice Hall Englewood Cliffs.

Rabiner, L. R., & Schafer, R. W. (2009). Theory and application of digital speech processing. *Preliminary Edition*.

Rajibul Islam, M., Shohel Sayeed, M., & Samraj, A. (2008). *Biometric template protection using watermarking with hidden password encryption.* Paper presented at the Information Technology, 2008. ITSim 2008. International Symposium on, 1-8.

Rec, I. (1988). G. 711: Pulse code modulation (PCM) of voice frequencies. *International Telecommunication Union, Geneva, 18*.

Rec, I. (1996). P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*.

Rekik, S., Guerchi, D., Selouani, S.-A., & Hamam, H. (2012). Speech steganography using wavelet and Fourier transforms. *EURASIP Journal on Audio, Speech, and Music Processing, 2012*(1), 1-14.

Reynolds, D. (2010). *An Overview of Automatic Speaker Recognition*. Paper presented at the JHU 2010 Workshop Summer School MIT Lincoln Laboratory.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech communication, 17*(1), 91-108.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(1), 19-41.

Roberts, C. (2007). Biometric attack vectors and defences. *Computers & Security, 26*(1), 14-25.

Sang, J., Liao, X., & Alam, M. (2006). Neural-network-based zero-watermark scheme for digital images. *Optical engineering, 45*(9), 097006-097006-097009.

Sarkar, G., & Saha, G. (2009). *Efficient pre-quantization techniques based on probability density for speaker recognition system.* Paper presented at the TENCON 2009-2009 IEEE Region 10 Conference, 1-6.

Satonaka, T. (2002). *Biometric watermark authentication with multiple verification rule.* Paper presented at the Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on, 597-606.

Schroeder, M. R., Atal, B. S., & Hall, J. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America, 66*(6), 1647-1652.

Seyed Omid Sadjadi, M. S., and Larry Heck. (2013). MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research: IEEE.

Shlien, S. (1997). The modulated lapped transform, its time-varying forms, and its applications to audio coding standards. *IEEE Transactions on Speech and Audio Processing, 5*(4), 359-366.

Simon, J. (2012). DataHash. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/31272-datahash/content/DataHash.m

Swanson, M. D., Zhu, B., Tewfik, A. H., & Boney, L. (1998). Robust audio watermarking using perceptual masking. *Signal Processing, 66*(3), 337-355.

Taal, C. H., Hendriks, R. C., & Heusdens, R. (2012). A low-complexity spectro-temporal distortion measure for audio processing applications. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(5), 1553-1564.

Takahashi, A., Nishimura, R., & Suzuki, Y. (2005). Multiple watermarks for stereo audio signals using phase-modulation techniques. *Signal Processing, IEEE Transactions on, 53*(2), 806-815.

Unoki, M., Imabeppu, K., Hamada, D., Haniu, A., & Miyauchi, R. (2011). Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics. *J. Information Hiding and Multimedia Signal Processing, 2*(1), 1-23.

Vallabha, G. K., & Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech communication, 38*(1), 141-160.

Vatsa, M., Singh, R., & Noore, A. (2009). Feature based RDWT watermarking for multimodal biometric system. *Image and Vision Computing, 27*(3), 293-304.

Verdu, S., & Han, T. (1994). A general formula for channel capacity. *Information Theory, IEEE Transactions on, 40*(4), 1147-1157.

Vielhauer, C., Scheidat, T., Lang, A., Schott, M., Dittmann, J., Basu, T., & Dutta, P. (2006). *Multimodal speaker authentication–evaluation of recognition performance of*

watermarked references. Paper presented at the Proceedings of the 2nd Workshop on Multimodal User Authentication (MMUA), Toulouse, France, 1-8.

Villalba, J., & Lleida, E. (2011). *Speaker verification performance degradation against spoofing and tampering attacks.* Paper presented at the FALA workshop, 131-134.

Wakita, H. (1976). Residual energy of linear prediction applied to vowel and speaker recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 24*(3), 270-271.

Wang, X., Qi, W., & Niu, P. (2007). A new adaptive digital audio watermarking based on support vector regression. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(8), 2270-2277.

Wei-Zhen, J. (2010). *Fragile audio watermarking algorithm based on SVD and DWT.* Paper presented at the Intelligent Computing and Integrated Systems (ICISS), 2010 International Conference on, 83-86.

William, S. (2006). *Cryptography and Network Security, 4/e*: Pearson Education India.

Woo, R. H., Park, A., & Hazen, T. J. (2006). *The MIT mobile device speaker verification corpus: data collection and preliminary experiments.* Paper presented at the Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The, 1-6.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: a survey. *Speech communication, 66*, 130-153.

Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E., . (2012). *A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case.* Paper presented at the Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC).

Xiang, S. (2011). Audio watermarking robust against D/A and A/D conversions. *EURASIP J. Adv. Sig. Proc., 2011*, 3.

Yan, B., & Guo, Y.-J. (2013). Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization. *Multimedia Tools and Applications, 67*(2), 383-405.

Yong-mei, C., Wen-qiang, G., & Hai-yang, D. (2013). An Audio Blind Watermarking Scheme Based on DWT-SVD. *Journal of Software (1796217X), 8*(7).

Zhe-Ming, L., Bin, Y., & Sheng-He, S. (2005). Watermarking combined with CELP speech coding for authentication. *IEICE Transactions on Information and systems, 88*(2), 330-334.

Zheng, N., & Adviser-Ching, P.-C. (2006). *Speaker recognition using complementary information from vocal source and vocal tract*. Ph. D. dissertation, The Chinese University of Hong Kong, People's Republic of China.