Doctoral Dissertation

Singular Spectrum Analysis-based Auditory Information Hiding with CNN-based Parameter Estimation for Speech Security

Kasorn GALAJIT

Supervisor: Professor Masashi UNOKI

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]
September 2021

Abstract

Speech signals are adopted in various forms and many social applications in a cyber-physical system (CPS), such as voice command, voice activation, and voice recognition. However, high-end speech editing software, such as voice conversion techniques and speech synthesis software, makes anyone easily fabricate and alter speech signals. These misused of this technology create risk in the security of speech technology and lead to social problems according to the increasing of unauthenticated speeches. These unauthenticated speeches can be used for criminal purposes such as theft or fraud in any systems in CPS. The attacks of unauthenticated speech signals, such as tampered speech, spoofed speech, and modified speech are considered an emerging threat. Thus, it is necessary to provide security of speech signals. Cryptography is a classical method that provides security by concealing speech signals from being tampered with and modified. However, cryptography does not detect tampering and modification in speech signals. Auditory information hiding (AIH) is one of the solutions to provide speech security by creating a secret channel and detecting tampering.

This research aims to provide security for speech signals in two objectives. The first objective is security in terms of protecting the genuineness of the speech signal. If attackers try to modify or change the speech signal, AIH can be used to protect its genuineness by tampering detection. One crucial property of information hiding is that the hidden information should be difficult to remove from the watermarked signal, and if there are attacks performed on the watermarked signal, the hidden information should reflect that change. The second objective is to protect the secret communication of the speech signal. AIH can be used to build the secret channel, and the transformation is used to secure the secret data on the secret channel.

Based on literature reviews, several information hiding techniques have been previously developed, and the singular spectrum analysis (SSA)-based AIH showed its strength in robustness due to the invariance of the singular spectrum. Moreover, SSA-based AIH could be designed to gain semi-fragile property (robust against non-malicious attacks but fragile to malicious attacks) by properly selecting part of the singular spectrum to be modified. The possibility of semi-fragile in SSA-based AIH motivates to construct a scheme for tampering detection. In addition, we deployed the convolutional neural network (CNN) method for parameter estimation instead of the differential evolution-based method adopted in the original SSA-based AIH.

For the first objective, the experimental results showed that the proposed scheme could locate tampered areas correctly, and it could also predict the types and degrees of tampering roughly. CNN-based parameter estimation could significantly reduce computational time, and the scheme is entirely blind because the estimation could be used to suggest the parameters in both embedding and extraction processes. However, the tampering detection accuracy needs to be improved since the proposed scheme is fragile to MP4 and robust to echo adding.

For the second objective, we cooperate transformation techniques with our SSA-based AIH to construct the secret and secured channel. The experimental results show that SSA-based AIH cooperated with Arnold transformation technique can provide the secret and secured channel. Only the authorized person with the correct key can access data at each level.

Index Terms: Singular spectrum analysis, SSA-based information hiding, CNN-based parameter estimation, tampering detection, speech security

Acknowledgements

My deepest gratitude goes to Prof. Masashi Unoki, Dr. Pakinee Aimmanee, and Dr. Jessada Karnjana, my supervisors at JAIST, SIIT, and NECTEC, respectively. I am grateful to my advisor, Prof. Masashi Unoki, for his tremendous guidance and support. Even he is so busy on the duty of Dean of information science school, he always gives time for his student to discuss and advice for a good direction for my research. Dr. Pakinee gives me advice and guidance in academic field and gives me a peaceful mind to rest from the stress of Ph.D. life, and encourages me with an extraordinary warm heart. My advisor from NECTEC, Dr. Jessada Karnjana, both a friend and my motivation to study, without him I was not thinking of being a Ph.D candidate.

I would like to extend my heartfelt gratitude to Prof. Masato Akagi whose always give valuable questions and comments on my laboratory meeting that helped me to improve my knowledge and logical thinking. I would also like to extend my sincere thanks to my friends and Acoustic and Information science Laboratory members. We have a great time of study together, discuss together, and comments on each other work. They are of great importance to me.

I appreciate the supporting grants from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program, Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233), and a Grant-in-Aid for Scientific Research (B) (No.17H01761).

Contents

Abstract **i**

Acknowledgements **iii**

List of Figures **vii**

List of Tables **ix**

List Of Symbols/Abbreviations **x**

1 Introduction **1**
1.1 State of the problem and research importance 1
1.2 Challenge 8
1.3 Motivation and Research Goal 9
1.4 Thesis Outline 11
1.5 Summary 12

2 Literature Review **14**
2.1 Auditory Information Hiding (AIH) 14
2.1.1 Main concept of auditory information hiding 14
2.1.2 AIH system 17
2.1.3 AIH techniques 17
2.1.4 AIH applications 21
2.2 Singular Spectrum Analysis 22
2.2.1 Stage 1: Decomposition 22
2.2.2 Stage 2: Reconstruction 24
2.3 Tampering Detection: state-of the art 25
2.3.1 Tampering definition 25
2.3.2 Tampering detection method 26

3 SSA-based AIH core structure for tampering detection 31
3.1 SSA-based AIH for tampering detection and its issues 31
3.2 Philosophy of this work 32
3.3 The core structure of SSA-based AIH for tampering detection 33
3.3.1 Embedding Process 34
3.3.2 Extraction Process 38
3.3.3 Tampering Detection 38
3.4 CNN-based Parameter Estimation 39
3.4.1 Training CNN 40
3.4.2 Generating High-Quality Dataset 41

4 Evaluation and Results 45
4.1 Dataset and Conditions 45
4.2 Sound Quality Evaluation 46
4.3 Semi-fragility Evaluation 47
4.4 Tampering Detection Ability 48
4.4.1 Tampering detection accuracy 52
4.5 Computational Time 53
4.6 Discussion 57
4.7 Summary 60

5 Application of Information Hiding 62
5.1 Statement of the problem 63
5.2 Background 63
5.2.1 Singular spectrum analysis-based Information hiding . 63
5.2.2 Arnold scrambling algorithm 64
5.3 Proposed method 65
5.3.1 The scheme for construct secret and secure channel . . 65
5.3.2 The scheme deployed encryption 65
5.3.3 Emitter side 66
5.3.4 Receiver side 69
5.4 Evaluations and results 71
5.4.1 Scheme used for building secret and secure channels evaluation 71
5.4.2 Scheme deployed encryption evaluation 72
5.5 Summary 74

6 Conclusion 76
6.1 Summary 76
6.2 Contribution 78
6.3 Future Work 79

Bibliography **81**

Publications **90**

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhorn International Institute of Technology, Thammasat University.

List of Figures

1.1 Application over internet and its the information flow between cyber space and connected physical devices. 2
1.2 Speech synthesis software: (a) WORLD and (b) STRAIGHT. 4
1.3 Cryptography method for security of speech signals. 5
1.4 Requirement of information hiding system. 7
1.5 Scope of speech security of interested. 9
1.6 Dissertation structure. 13

2.1 Information hiding classification proposed by FA. Peticolas et al. 16
2.2 AIH system. 18
2.3 The basic SSA. 23
2.4 A tampering detection applied in ASV system. 27
2.5 A tampering detection system. 28
2.6 A tampering detection system using information hiding. 30

3.1 Singular spectrum of one frame. 32
3.2 Singular spectrum of the four different speech segments. 33
3.3 Proposed framework: embedding process (left) and extraction process with tampering detection (right). 34
3.4 Example of the part of a singular spectrum $\{\sqrt{\lambda_p}, \sqrt{\lambda_{p+1}}, ..., \sqrt{\lambda_q}\}$: (a) selected part of singular spectrum without embedding, (b) watermark bit 1 is embedded, and (c) watermark bit 0 is embedded. The red line shows the threshold level $\gamma \cdot \sqrt{\lambda_0}$, and the blue dashed line connects from $\sqrt{\lambda_p}$ to $\sqrt{\lambda_q}$. 37
3.5 Decoding hidden watermark bit: if most of singular values (circle) that are under threshold level $\gamma \cdot \sqrt{\lambda_0}$ are above blue dashed line, extracted watermark bit is 1, but if most of singular values (asterisks) that are below threshold level $\gamma \cdot \sqrt{\lambda_0}$ are under blue dashed line, extracted watermark bit is 0. . . . 39

3.6 Structure of two CNNs: (a) CNN used to estimate embedded strength parameters and (b) CNN used to estimate parameter of γ. 41
3.7 DE optimizer used to create dataset 43
3.8 Framework for creating training dataset. 44

4.1 Comparison of watermark image between original image (a) and reconstructed images after performing following signal-processing operations: (b) no attacks, (c) MP3, (d) G.711, (e) G.726, (f) MP4, (g) PSH -20%, (h) PSH $+20\%$, (i) SCH $+4\%$, (j) SCH -4%, (k) BPF, (l) PSH -10%, (m) PSH $+10\%$, (n) AWGN (40 dB), (o) echo (100 ms), (p) replace (1/3), (q) PSH -4%, (r) PSH $+4\%$, (s) AWGN (15 dB), (t) echo (20 ms, and (u) replace (1/2). 51
4.2 Comparison of extracted watermark-image: (a) no attacks and (b) second half of speech signal substituted by synthesized speech signal. 51
4.3 RMSE of γ, μ, and σ from DE-based parameter estimation and CNN-based parameter estimation. 55
4.4 Example of singular spectrum of embedded frame. "$\diamond$" denotes original singular spectrum, "$*$" denotes modified singular spectrum where parameters are obtained from CNN-based method, and "$\circ$" denotes singular spectrum where parameters are obtained from DE-based method, red solid line denotes γ threshold, and dashed line denotes a straight line connected the first and last singular value to be modified. 56

5.1 Secret and secure channel: Emitter (left), and receiver (right). 65
5.2 Scheme deployed encryption: Emitter (left), and receiver (right). 66
5.3 Emitter side. 67
5.4 Receiver side. 70

List of Tables

4.1 Sound-quality evaluations: proposed scheme vs. other methods. 47
4.2 BER (%): proposed scheme vs. other methods. 48
4.3 An accuracy of tampering detection of the proposed method . 52
4.4 An accuracy of tampering detection after defined threshold of attacks 53
4.5 Comparison of computational times for of parameter estimation of the method based on differential evolution and the method based on CNN. 54
4.6 Comparison of robustness and inaudibility of scheme when automatic parameterization is based on differential evolution and when it is based on CNN. 54
4.7 Eight cost functions studied in our investigation. 58
4.8 Evaluations of robustness and inaudibility when different cost functions were deployed. 59

5.1 Comparison of imperceptible properties between proposed and other methods 72
5.2 Comparison of correlation coefficient and SNR (in dB) between original speech (ori), encrypted speech (enc), and decrypted speech (dec) for the proposed method and other encryption methods. Note that NA is not applicable data . . . 73
5.3 Key sensitivity and BER 74

List Of Symbols/Abbreviations

AIH	Auditory information hiding
ASV	Automatic speaker verification
AWGN	Adding white Gaussian-noise
BER	Bit-error rate
BPF	Band-pass filtering
CD	Cochlear delay
CELP	Code-excited linear prediction
CNN	Convolutional neural network
CPS	Cyber-physical system
CPU	Central processing unit
DCMP	Digital circuit multiplication equipment
DCT	Discrete cosine transform
DE	Differential evolution
DFT	Discrete Fourier transform
DWT	Discrete wavelet transform
dB	Decibel
F0	Fundamental frequency
FAR	False acceptance rate
FE	Formant enhancement
FFT	Fast Fourier transform
FN	False negative
FP	False positive
FRR	False rejected rate
GMM	Gaussian mixture models
HAS	Human auditory system
HMM	Hidden Markov Model
ITU	International Telecommunication Union
LSB	Least-significant-bit
LSD	Log-spectral distance
MGD	Modified group delay
MOS	Mean opinion score

NSCR	Number of sample change rates
PESQ	Perceptual evaluation of speech quality
PSH	Pitch shifting
QIM	Quantization index modulation
ReLU	Rectified linear unit
RMSE	Root-mean-square error
RSA	Rivest-Shamir-Adleman
SCH	Speed changing
SDR	Signal-to-distortion ratio
SNR	Signal-to-noise ratio
SS	Spread spectrum
SSA	Singular spectrum analysis
SVD	Singular value decomposition
UBM	Universal background model

Chapter 1

Introduction

This chapter states the problem that we want to solve and why this problem is worth solving. In this study, we want to solve the security problem in the speech signal by using the information hiding method. Initially, we want to clarify the meaning of security we focus on in this study first. The security in our study considers two aspects. The first aspect is security in terms of protecting the genuineness of the speech signal. If attackers try to modify or change the speech signal, we can use our proposed information hiding method to protect speech genuineness, i.e., to detect tampering in a speech signal. The second aspect is to protect the secret communication on the speech signal. We applied our proposed information hiding method with the transformation technique to build a secret and secured channel.

Note that in this study, we show speech security in two scenarios: first, to detect tampering, and second, to build the secret and secured channel. These two scenarios are not related to each other that is no tampering detection on the second scenario, but both scenarios apply the same core structure of our proposed information hiding method.

After we state the problem and explain why this problem is worth solving, we will then explain the motivation why we think the information hiding can solve this problem. Then, the challenge of solving this problem will be discussed. Finally, we will show the goal to be achieved in this study.

1.1 State of the problem and research importance

The rapid growth of the Internet has positively impacted societies and communities in many ways. There are enormous services and applications through the Internet that make human life more accessible and convenient.

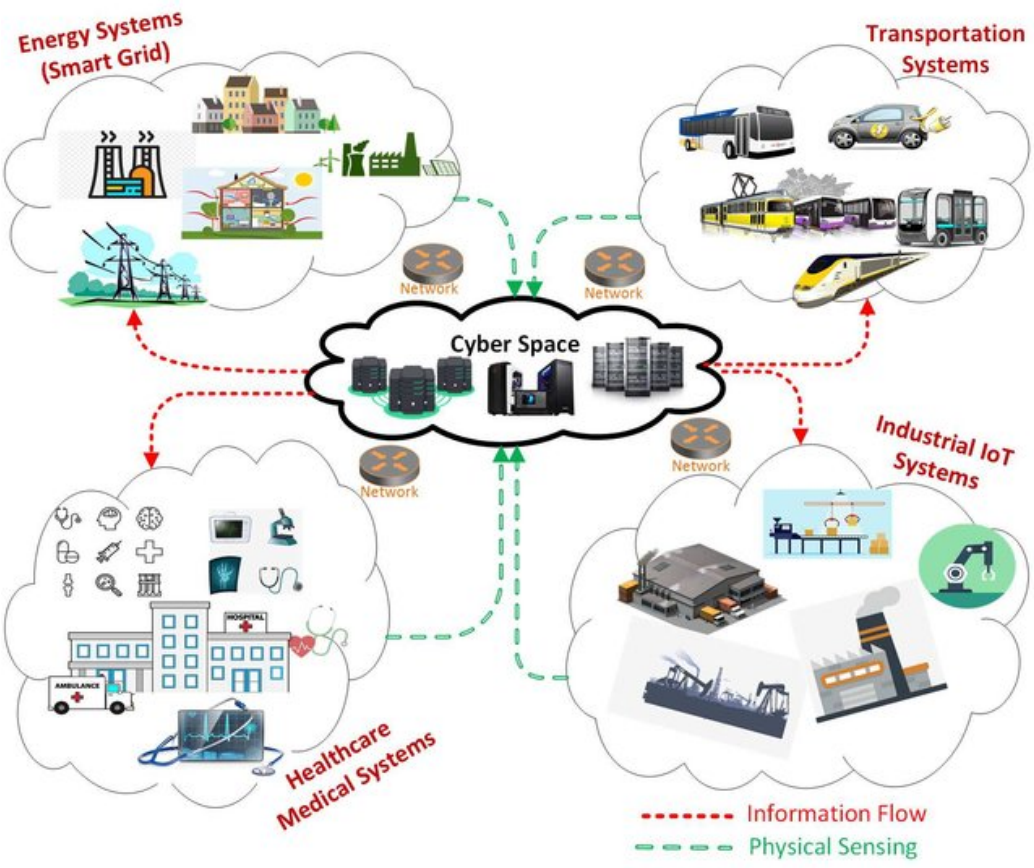


Figure 1.1: Application over internet and its the information flow between cyber space and connected physical devices.

Living in an information society, we cannot avoid the fact that there is enormous data exchange through the Cyber-physical system. Figure 1.1 showed applications over the Internet and its the information flow between cyberspace and connected physical devices [1]. The various type of data was exchanged through the network. Those can be personal data or data sent from a mobile, a computer, or specific electronic devices. Moreover, those data can be in various forms such as raw files, digital files, videos, audio, and images.

In this study, we focus on the data in the form of the speech signal. Speech signals are adopted in various forms and many social applications in an information society, such as voice command, voice activation, and voice recognition. With the fast development of advanced digital technologies, speech services and speech applications are increasing in number. For example, A text-to-speech program and speech synthesis program can be helpful

for disabled people to communicate in daily life [2, 3, 4]. A program that used speech synthesis software called "Speech Morphing" can create emotionally intelligent voices for more natural conversations to respond to customers [5]. A Voice recognition system applied in internet banking makes people access their accounts using their voice on mobile without facing traffic problems on the road.

As we know, the increase in usage, the more possibility of being threats, and the network-based threats have become more sophisticated. Since speech contains vital and essential information, it attracts the attackers to steal or modify speech to fault the service system. Therefore, when speech signals are exchanged or transmitted, they need to be protected from modification, manipulation, forgery, and theft.

Moreover, high-end speech editing software, such as voice conversion techniques and speech synthesis software, makes anyone easily fabricate and alter speech signals. For example, speech synthesis software such as STRAIGHT, WORLD can be used to modify voice on its timbre, pitch, speed, and other attributes flexibly [6, 7]. Figure 1.2 showed synthesis software such as STRAIGHT, WORLD that used to modified speech signal. These advanced digital tools and technologies make people can easily modify or alter speech without prior knowledge. Therefore, these misused of this technology create risk in the security of speech technology and lead to social problems according to the increasing of unauthenticated speeches. These unauthenticated speeches can be used for criminal purposes such as theft or fraud in any system. The attacks of unauthenticated speech signals, such as tampered speech and manipulated speech, are considered an emerging threat. Therefore, it is necessary to provide security for speech signals.

Let start to see the importance of the security problem in the speech signal from the following example. The first example is considered about speech recording to be used as evidence in the court. As we mentioned earlier, there are massive tools to edit or tamper speech. People may modify the content in the recording by cutting or replacing the conversation to cheat on justice. Therefore the voice recording used to prove the case must comply with the following, Firstly, the conversation in the recording is relevant to the case. Secondly, there is an identification of the voice. Thirdly, the recording is the genuine one [8]. To solve this problem, we need a tools to check whether the recording is a genuine one or tampered one.

The second example is concerning speech using in the voice recognition system or speaker verification system. Since voice contains personal information and it is claimed that the voice of each person is unique. Voice interfaces have become more popular, and many voice service systems integrated recognition capabilities. So the system with recognition capabilities

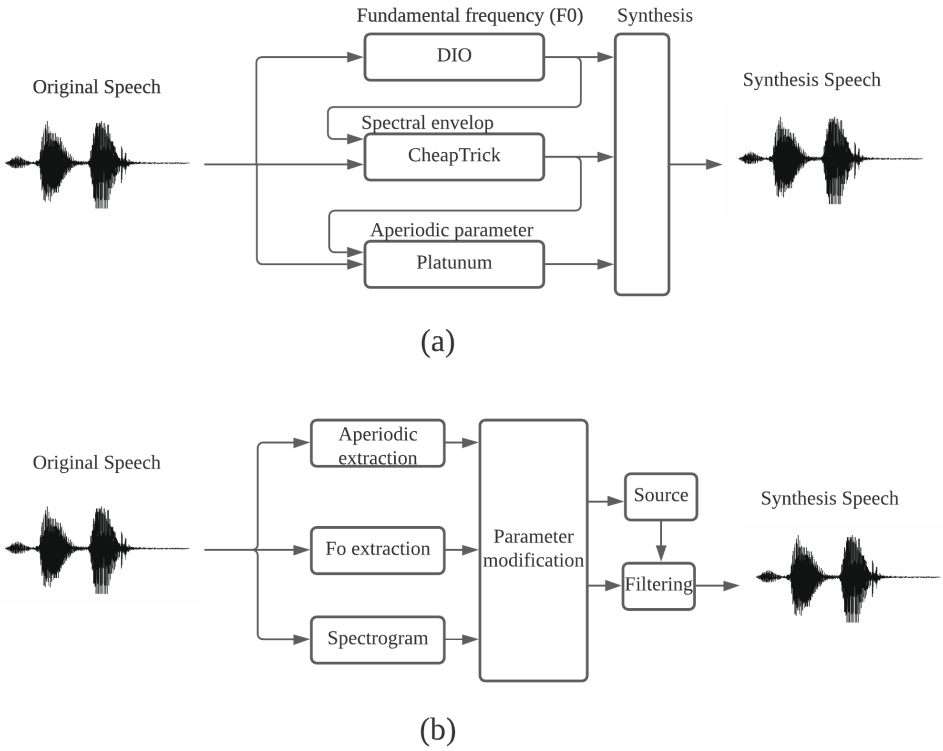


Figure 1.2: Speech synthesis software: (a) WORLD and (b) STRAIGHT.

can understand what has been said (refer to speech recognition) and who has said it (refer to speaker recognition). Nowadays, this type of service is available in many commercial products. For example, the well-known one is Google home, and Siri on Apple [9]. In the financial sector, the well-known bank like HSBC and Lloyds Bank also deployed the authentication service to access bank account on mobile phone [10, 11]. However, in the past few years, there was a big shock that the voice recognition system of HSBC was faulted from the mimic voice [10]. A mimic or spoofed voice used to fault the system can obtain from using a voice conversion or voice synthesis program by tuning some characteristic of voice to close to the target. The attacks on the financial sector are sensitive since involving money. The accuracy of the speaker verification system can be improved if there is a countermeasure for pre-processing the incoming speech before feeding it to the speaker verification system. This countermeasure determines the genuineness of incoming speech and eliminates those modified and tampered speech.

The last example concerns the situation that two people want to commu-

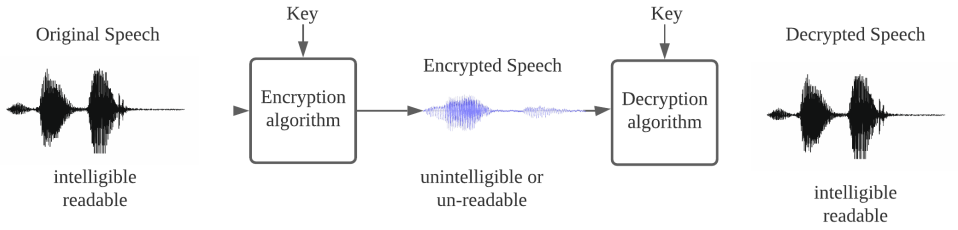


Figure 1.3: Cryptography method for security of speech signals.

nicate secretly, but they do not want others to know that they are privately communicating. In some companies, they monitor the email, song, recording that send through the network. Their policy does not allow the employees to encrypt the message, including a song or sound recording, since they cannot monitor those encrypted messages.

It seems that to resolve these three examples, we have to propose a different solution for each problem. If we consider the first and the second example, the cryptography can be used the solve the first two example. Cryptography is a classical method that solves the problem by concealing speech signals to prevent them from being stolen and modified. It solves the first example by if speech is encrypted then you can not cut, replace or edit the content. Also its characteristic cannot be stolen to do synthesis or conversion to fault system in example two. However, once the content has been decrypted, such that cryptography does not protect speech signals anymore [12]. The modification and the alter of speech after decrypted cannot be tracked. When consider on the third example, the cryptography cannot solve this problem since it can only conceal the message but it cannot hide the face that two person try to communicate secretly. Some countries have some restrictions on encryption technologies, such as law enforcement access to encrypted data [13, 14]. This restriction makes it difficult for cryptography method usage. Figure 1.3 showed cryptography method for security of speech signals.

Besides cryptography, some traditional techniques are adopted to solve the security problem in speech signals, such as scrambling and transformation. Scrambling is performed by segmenting speech signals into small elements and shuffle them, while transformation is scrambling speech signals in another domain such as time, frequency domain, or both domain. Scrambling and transformation are a method to change the intelligible speech signals into unintelligible signals. The concept of scrambling and transforming is similar to cryptography, as well as their disadvantage is similar. One crucial

disadvantage is that the unintelligible of signals can cause the attention of attackers because the data stream of an encrypted signal, scrambled signal, and transformed signal are random and meaningless gibberish. Once the gibberish is intercepted, then the attackers found the target to attack. All cryptography, scrambling, and transformation mainly provide the security to protect speech from being tampered with, but none of them could be used for tampering detection.

Auditory information hiding (AIH) can be a potential solution to solve all the mentioned three example [15, 16]. Auditory information hiding is a method that is embedding hidden information unnoticeable into a speech signal. In other words, the listeners are not even aware of the existence of hidden information. Once the speech signal is altered or modified, the changing of hidden information can track the alter and modification. Thus if we use AIH for tampering detection it can solve the first and the second example. To solve the third example, from the property that the listeners are not even aware of the existence of hidden information so we can use the hidden information as a secret channel in speech signal.

As we mentioned, we want to solve the security problem for the speech signal, and our definition for speech security refers to two aspects: firstly, to detect tampering in a speech signal and secondly, to build a secret and secured channel. The auditory information hiding will be used to provide speech security. We will then consider the requirement in order to build the auditory information hiding system.

Auditory information hiding requirement

Typically, there are five requirements for auditory information hiding [17, 18]. Figure 1.4 illustrated the requirement of information hiding system.

1. *Inaudibility.* Inaudibility is a principal requirement of auditory information hiding. The embedding procedure should not make any degradation in the host signal in terms of sound quality, and the human auditory system should not hear the watermark embedded into the host signal.

2. *Robustness.* The robustness refers to high extraction precision, i.e., the hidden information should be correctly extracted even when attacks are performed on the watermarked signal. This property is used to confirm hidden information has been maintained while being transmitted over the channel communication or stored in the system. Also, the

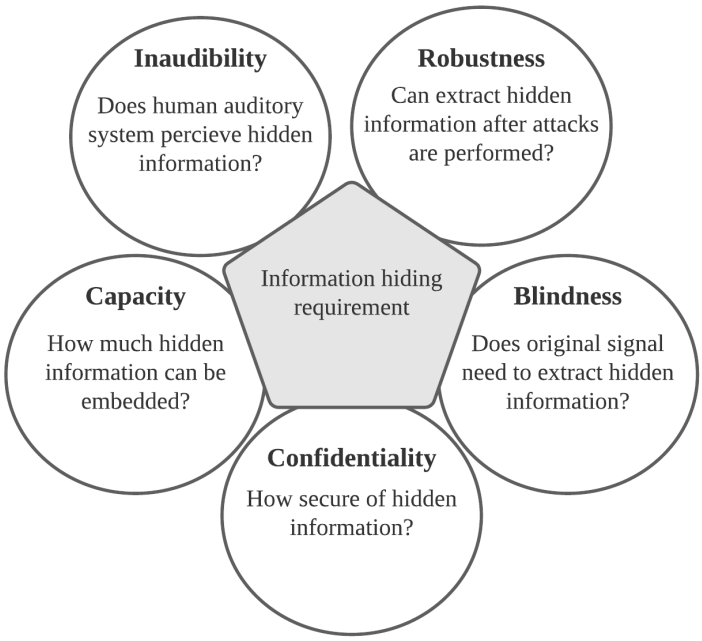


Figure 1.4: Requirement of information hiding system.

hidden message should survive the application of any speech processing techniques.

3. *Blindness.* Blindness refers to an ability to extract hidden information correctly without the host signal required.

4. *Confidentiality.* A property of concealing hidden information that is the hidden information must be secured.

5. *Capacity.* The capacity refers to the maximum quantity of the hidden information embedded into the host signal.

This work focused on three properties: inaudibility, robustness and blindness. Since our first objective is to construct the information hiding to detect tampering, one important property that we are looking for is that the scheme must be fragile to the attack but robust to normal signal processing. This property refers as semi-fragile property. Therefore, the scheme not only satisfy the requirement of basic information hiding but also must satisfy the semi-fragility.

1.2 Challenge

The challenge of to design the information hiding are as follows.

1. *Sensitivity of human auditory system.* By nature, the human auditory system is very sensitive, i.e., we can hear a sound wave with tiny pressure fluctuation [19], and adding hidden information into a host signal is as same as adding noise to a host signal. Therefore, this fact leads to the first challenge that making the information hiding to be imperceptible is a difficult task itself.

2. *Requirement balancing.* Typically, AIH requirements conflict with each other. For example, some techniques are good at robustness, but it is not blind [20]. The high capacity of hidden data comes with the cost of low speech quality which the listeners may suspect of any hidden information on the hearing speech [21]. Therefore, in addition to proposing new techniques, researchers in this field have to focus on compromising these conflicts, which has proved difficult. Thus, the second challenge is to solve the problem of conflicting requirements. To demonstrate the conflicting. Let consider especially the robustness property. Robustness will ensure that the embedding procedure should strictly attack the watermark with the host signal that cannot be easily removed. Thus, this research applies the information hiding for tampering because if the content or composition of the host signal changed, the hidden information should reflect that there are some modifications to the host signal composition. This fact leads to the challenge of making the embedding algorithm robust against speech signal operations and meanwhile fragile to reflecting the changing due to tampering and modification.

3. *Attacker creativity.* To create the system, we cannot predict about the motivation of attacker to attacks, remove, or destroy the hidden information. Thus, it is not easy to handle with possible all attacks.

4. *Complexity.* Some information hiding schemes can achieve good performance, but it might consume time on computation. Some use low computational time, but the schemes are large designs. The good AIH should be practical to implement.

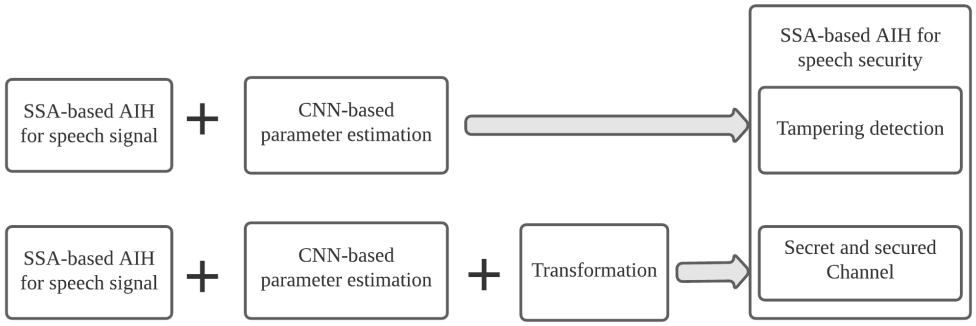


Figure 1.5: Scope of speech security of interested.

1.3 Motivation and Research Goal

In this work, we want to propose the information hiding method to solve the security problem in the speech signal. The security in our study considers two aspects. The first aspect is security in terms of protecting the genuineness of speech signals. The hidden information will be embedded into host speech, the hidden information will reflect the changing or modification of host speech. So it can prove the genuineness of the speech i.e., to detect tampering in a speech signal. The second aspect is to protect the secret communication on the speech signal. We applied our proposed information hiding method with transformation technique to build a secret and secured channel. Therefore, it can be said that we aim to build information hiding for two objectives: tampering detection and provide a secret and secured channel. Figure 1.5 illustrated the scope of speech security of interested.

Since our first objective is to construct the information hiding to detect tampering, one important property that we are looking for is that the scheme must be fragile to the attack but robust to normal signal processing. During reviewing the literature, we found that a method based on the least significant bit is fragile to attack but not robust to various signal operations [22]. Some have can fragile to the attacks but robust to only few signal operations [23].

Lastly, the method based on singular spectrum analysis (SSA) is selected [24]. This method uses the techniques by slightly changing the singular value of a matrix representing a host signal. This technique derives due to the invariance of singular values [25]. This SSA-based method has one attractive property: the different embedded areas have a different fragile degree. The embedding area on low order singular value increases robustness while embedding area on high order singular value increases fragility. From this property, we hypothesis that this method can be made semi-fragile method

for detecting tampering in a speech signal. This method initially applies to audio signals, not speech signals, and it works well with an audio signal. If The original SSA-based method can also apply to the speech signal, the scheme can have the advantage over the method that can apply only speech or audio signals. However, we found that the original SSA-based method has a critical problem since the parameter need in the embedding and extraction process requires a considerable time to estimate the embedding parameter. Therefore to achieve the first objectives to construct the information hiding for tampering detection, we have a few sub-goals step by step as follows.

1. To verify that SSA-based AIH applied on speech signal can keep the SSA-based technique advantage as it has done on an audio signal.

2. After we verify that the SSA-based method can be applied to speech signals, the next step we construct the scheme can be used to detect the tampering in the host speech. From the property of the SSA-based method: the different embedded areas have a different fragile degree. We have to select the area to be embedded and achieve semi-fragility property. Moreover, these embedding parameters can be adjusted to obtain a better performance.

3. Since the original SSA-based method uses a differential evolution (DE) to provide embedding parameters and has a problem with computational time. We offer a novel method to obtain the embedding parameter by suggesting of convolution neural network (CNN). However, CNN needs supervising to estimate this parameter. Since the DE gave the promising result, we designed to use parameters suggested by DE to train CNN. In this step, we must design a DE optimizer and verify that parameter obtained from the DE-based method give good performance and reasonable to use for CNN training.

4. Next, to verify that CNN-based parameter estimation can reduce the computational time and still can keep the requirement balancing as the original SSA-based AIH.

To reach all mentioned sub-goals, we expect to obtain the SSA-based information hiding that can be used for tampering detection, and the computational time for parameter estimation is reduced.

The second objective is motivated by the hidden information is not attract the attention of listeners. Thus, the hidden information can be used as a secret channel. The purpose of the second objective is to protect the secret communication on the speech signal. There is only one sub-goal on this stage.

1. To verify that the SSA-based AIH scheme can cooperate transformation to provide security of speech signal. Note that achieving this sub-goal, the concept of SSA-based information hiding for speech signal and CNN-based parameter estimation is also applied, and note that tampering ability is not considered in this sub-goal.

1.4 Thesis Outline

The organization of this dissertation is shown in Figure 1.6. Beside this introduction chapter, the rest of this dissertation consists of five chapters and is organized as follows.

Chapter 2 introduces the background knowledge of speech information hiding, the application that the speech information hiding applied for, and some conventional and previously advanced techniques. Since the proposed framework is based on singular spectrum analysis; therefore, the singular spectrum analysis is reviewed and analyzed carefully to recognize their strength and weakness. In this research, we apply the speech information for a specific purpose, i.e., tampering detection. Therefore, in this chapter, the state of the art for tampering and manipulating detection are reviewed to illustrates how different the classic method and the proposed method are. Lastly, the information hiding schemes for the purpose of tampering detection are reviewed.

Chapter 3 proposed the core structure framework that applied singular spectrum analysis into speech information hiding for tampering detection. The secret information, called watermark signal, embedded into speech signal is used for tampering and detection. The parameter finding suggests whether which part of the singular spectrum to hide the secret information. This chapter shows how good parameter estimation can help the scheme balance between inaudibility and robustness as well as be fragility to detect the modification on speech signal. Lastly, we proposed the SSA-based AIH method with parameter estimation using CNN.

Chapter 4 reports the evaluation of the core structure with CNN-based parameter estimation. This chapter explains the measurement and criteria to evaluate the performance of the scheme. The scheme will be evaluated in three aspects: the sound quality of the watermarked signal compared with the original signal, the scheme's robustness against many signal operations and attacks, and the ability to detect tampering and tampering detection accuracy. In addition, the computational time of the scheme that used CNN-based parameter estimation and the one used DE-based parameter estimation are compared. Moreover, we also discuss the effectiveness of CNN-based

parameter estimation and tampering detection.

Chapter 5 describe the SSA-based AIH for the second aspect and its experimental results. The SSA-based AIH scheme for protecting the secret communication on the speech signal is described. The SSA-based AIH scheme cooperates transformation to provide security of speech signal. Arnold transformation is deployed to provide the secret and secured channel. Lastly, the evaluation and result are provided.

Chapter 6 summarizes this research work and emphasizes the contribution of this work to the related research field. Since there is nothing entirely perfect, this chapter also discusses the room for improvement and future work.

1.5 Summary

The innovative points of this research can be concluded as follows: This research proposed the SSA-based AIH for tampering detection by incorporates CNN for parameters estimation. Moreover, to accommodate the security of speech signals, our SSA-based AIH is deployed transformation techniques to afford a secret and secured channel on speech signal.

In summary, this introduction chapter started with the problem statement, the problem we want to solve. Then we show how that problem and issues are essential for solving and why it is worth solving, including challenging to solve the problem. Next, the motivation and goal of this dissertation are identified. Finally, the structure of this dissertation is illustrated.

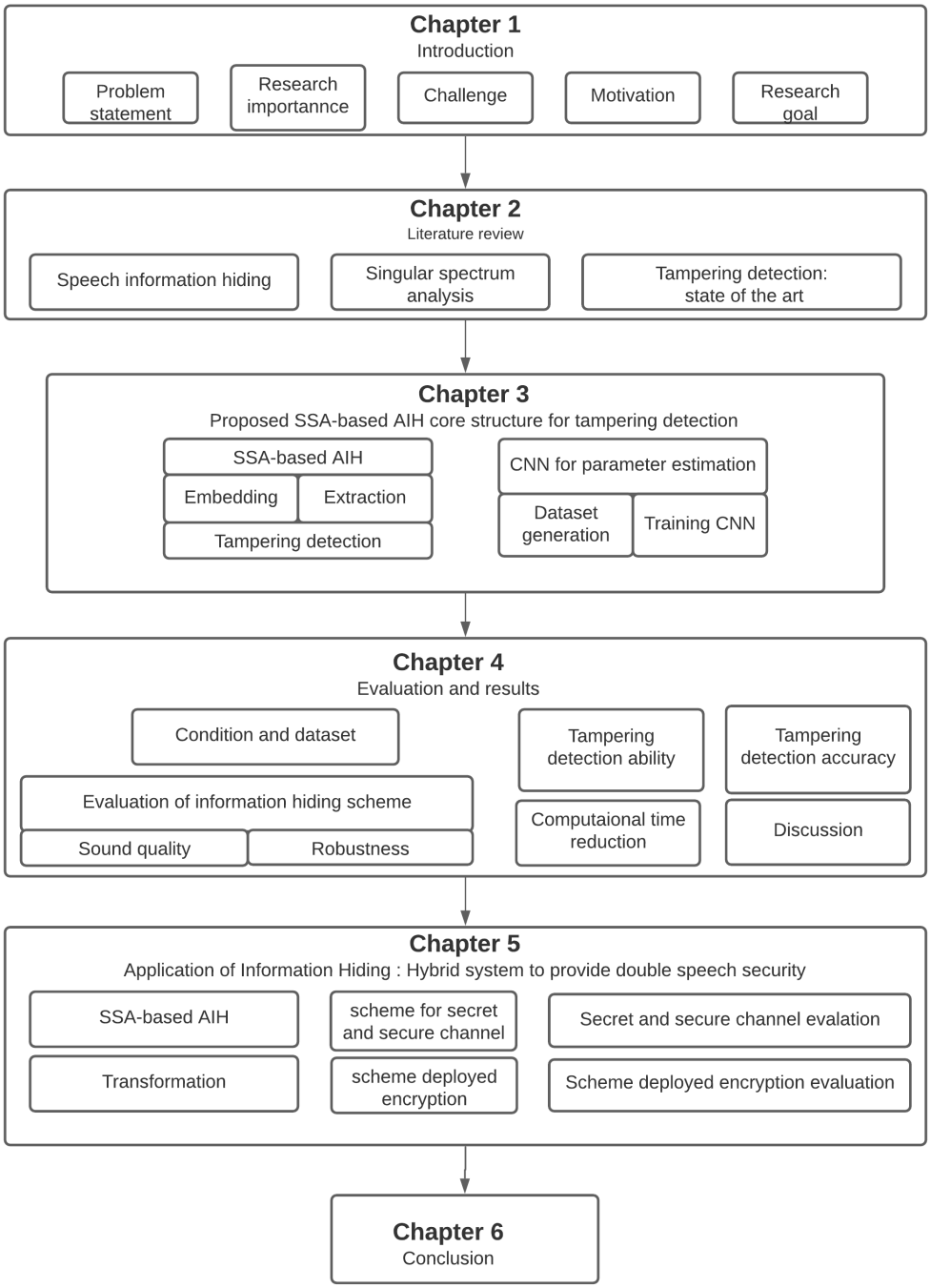


Figure 1.6: Dissertation structure.

Chapter 2

Literature Review

This chapter introduces the background knowledge and literature of auditory information hiding, their applications, and some conventional and previously advanced techniques. Since the proposed framework is based on singular spectrum analysis; therefore, the basic principle about singular spectrum analysis is reviewed and analyzed carefully to recognize their strength and weakness. In this research, we apply the speech information for a specific purpose, i.e., tampering detection. Therefore, in this chapter, the state of the art for tampering detection is reviewed to illustrates how different between the classic method and the proposed method.

2.1 Auditory Information Hiding (AIH)

This section provides an overview of auditory information hiding. We clearly define what auditory information is, what the purpose of information hiding. Besides, we also show information hiding schemes have been constructed with several techniques.

2.1.1 Main concept of auditory information hiding

The principal idea of information hiding originates from old-fashioned steganography, which aims at hiding essential messages into other information, and hidden information should not cause attention. Information hiding technology can be applied in various fields because both carrier and hidden information can be any kind of media. For example, hiding digital signatures as an image into printing image files, hiding an image into audio or speech signal, or hiding voice into an image file. Therefore information hiding is a technology that embeds the confidential information, namely

watermark signal, as a secret message into host information, called *public information* to obtain the carrier contained hidden information, called *watermarked signal*. The crucial point for information hiding techniques is that the hidden information embedded in the carrier should be entirely transparent. It should not attract attention or cause suspicion. For example, if an image-one is hidden into an image-two, we consider an image-one as secret information and consider an image-two as a carrier. After the embedding process, the audience should see only an image-two and do not suspect that an image-one exists in what they have seen. This concept is the same as biological camouflage that the animals try to adapt themself to look like their environment. Thus, information hiding refers to a method of concealment by means of disguise. The advantage of information hiding is that when the hidden information is transparent, then the chance or the possible risk of being attacked is minimized.

In this work, we focus on hiding the secret information into a speech signal. The difficulty is that the human perception system is susceptible. It can easily distinguish in the small changes of sound. Therefore, embedding the secret information into speech signals and making people unaware of their existence is a big challenge. We can imply that the changes in a characteristic of a carrier or host information should be minimizing so the listeners cannot hear the hidden information. In this circumstance, the watermark or secret information can be of any kind, and the host signal is a speech signal. The most significant advantage of information hiding is that only two authorized parties can access the hidden information while general audiences cannot sense the existence of a hidden message.

Since information hiding inherited the concept of classical steganography thus, the question of the difference between information hiding and steganography is always asked. In addition, the secret message is called a watermark, so that sometimes we call the method the watermarking method. Therefore the terminology: the information hiding, watermarking, and steganography confuses people. Strictly speaking, the words information hiding, watermarking, and steganography are different but closely related. Both watermarking and steganography are information hiding but with different aspects. The classifications of them also vary depending upon which criteria are used to classify. Figure 2.1 depicts one example of the information hiding classification proposed by FA. Peticolas et al. [26]. The information hiding is divided into several perspectives as described follows.

The first category, *covert channel* is a secret channel to send or receive the secret message. It does not use the typical channel to communicate, but it only deploys the information hiding technology to build the secret channel on a typical channel and prohibit access from unauthorized parties. The

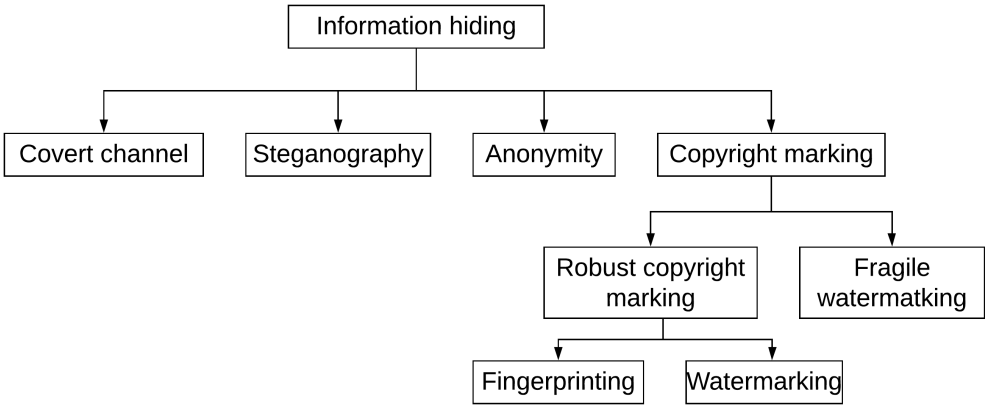


Figure 2.1: Information hiding classification proposed by FA. Peticolas et al.

covert channel is difficult to employ in practice because the advanced security technology can easily find out the covert channel. Once its existence is found, it cannot be secret anymore, and the entire information hiding design will be destroyed. The second category, *steganography* refers to a method that conceals one information within another information. The concealed one is called the secret message, and the carrier is the public information or the cover message. The concealed message should not cause any attention from observers or listeners. In some classifications, steganography and the covert channel are defined as the same principle method. Primarily steganography is used in one-to-one communication, and its main required property is that concealed message is challenging to detect. The third category, *Anonymity* refers to the anonymous communication mechanism. Since there are many activities on Internet, so sometimes private information needs to be protected. Anonymity describes the actions or activities on the Internet where the personal identity of that acting is unknown. However, attackers can easily find out personal information and attack that information. The final category called *copyright marking* has developed information hiding for copyright concerns. This category is sometimes called *watermarking* . It includes digital watermarks (both robust and fragile methods) and digital fingerprints. The watermarking method is used in one-to-many communications. The main difference between steganography is that steganography focus on concealing its existence, while watermarking gives more importance to making the hidden message difficult to removed or cannot replace easily. The steganography method does not provide robustness against removing or modification as watermarking does. Even information hiding is divided into four categories, but there are still overlapped depending upon the aspect we

examined.

Let us consider the scope of this research, and we intend to embed the hidden into the speech signal for tampering detection. Therefore, we can say that our hidden information should be fragile to the attacks but robust against common speech processing. This method is in between fragile and robust watermarking, so we define the terminology *semi-fragile* to refer to the property that is fragile to the attacks but robust against common speech processing. Besides, the terms information hiding and watermarking are interchangeable; thus, we also use information hiding and watermarking as a synonym of each other.

2.1.2 AIH system

An AIH system consists of two main processes: embedding and extraction, as presented in Fig. 2.2. If we consider in communication viewpoint, the embedding process will take place on the sender side while the extraction process is on the receiver side. The embedding process can be considered a function that a host signal and the confidential information called watermark as its input and returns a watermarked signal. Given the host signal A and the hidden information w, the watermarked signal A^* can be expressed mathematically by the following equation

$$A^* = A + f(A; w) \tag{2.1}$$

where, the function f is an embedding function.

The extraction process, sometimes called detection, extracts the hidden information $\hat{w}$ from the watermarked signal A^*. The process can be expressed mathematically by the equation.

$$\hat{w} = g(A^*; c(A)) \tag{2.2}$$

where the function g is an extracting function and $c(A)$ is a function representing some information that depends on the host signal A. If there is no such information (c(A)=0), the extraction process is called blind detection; otherwise, the non-blind detection.

2.1.3 AIH techniques

Since working online or communication through the internet increases significantly, digital security communication, copyright protection, anti-fake, and data integrity are more concerned. Information hiding techniques are being developed and have been more widely used in recent years. There are

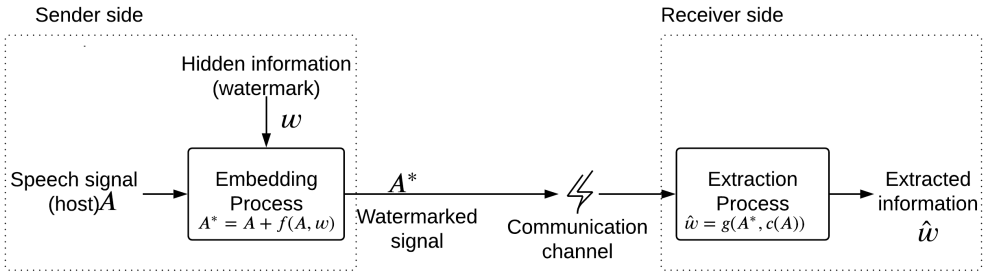


Figure 2.2: AIH system.

many ways to classify AIH techniques since there are several properties that we can use to define the algorithm. In other words, classification depend upon a set of criteria used for grouping. Many methods are based on the human auditory system (HAS), and several commonly used approaches related to a signal property. Therefore, we easily classify the audio watermarking techniques into four categories [27].

- Time domain hiding
- Echo hiding
- Transform domain
- Human auditory-based method

However, this classification is inconsistent because many methods fall into more than one category and mix across more than one technique. For example, the method based on the low-frequency amplitude modification [28] embeds information in the time domain and is based on the human auditory system, so it has both concepts of first and the last categories. There is a method based on the fast Fourier transform (FFT) amplitude interpolation, and the human auditory system [29]. Thus it has a concept of the third and the fourth categories. In this section, we try not to group the existing AIH techniques to avoid such confusion. The following subsections briefly explain some AIH techniques.

Time domain method

The time-domain information hiding is easy to implement and straightforward solution because it directly modifies the speech sample. A well-known example of time-domain hiding is the least-significant-bit (LSB) replacement-based method. It usually replaces the least significant bit of the host signal

with a binary bit of watermark signal [22]. Since the modification part is a minimum weighting value, the inaudibility can be ensured because the LSB hardly affects human perception. The advantage of this method can be seen clearly in two main points: it needs less computational time because it is not complex, and the second advantage is that it has a large embedding capacity. The significant disadvantage of the technique is not resistance to manipulations. In other words, the hidden information can be easily damaged by various signal-processing attacks. E.Erçelebi et al. attempted to increase the robustness of this technique using multi-bits generated by pseudorandom sequences embedded into the host signals [30]. However, increasing robustness in the time domain watermark is not easy to achieve due to host signal distortion.

Echo hiding

Echo hiding is sometimes categorized as a sub-domain of the time-domain method [31]. This method is based on the fact that if the weak signal appears after the strong signal quickly, then the human auditory system cannot detect that weak signal [32]. This phenomenon is called masking effects, and it is utilized to construct an information hiding scheme by introducing an echo signal to hide the secret information. An attenuated echo is added to the host signal in the embedding process and performs the watermark extraction in the extraction process with cepstrum analysis. The advantage of the echo hiding method is excellent inaudibility. The disadvantage is that the embedding process is signal-dependent. Consequently, it is difficult to make a blind information hiding scheme.

Transform domain method

In the third category, the watermark is embedded in a specific transform domain. The various transforms include frequency transforms such as fast Fourier transform (FFT) [29, 33], discrete cosine transform (DCT) [34, 35], and time–frequency transform such as discrete wavelet transform (DWT) [34, 36]. Transform domain watermarking applies a specific transform to the data block of the signal and then hiding the watermark into the transformed data block [37].

For all transform domain methods, they consist of two more steps compare with the time-domain method. The host signal $x(n)$ is forward transformed into the selection domain as $X(k)$ before watermark embedding. The signal $Y(k)$ obtain from embedding the watermark signal to $X(k)$, and then take an inverse transform of $Y(k)$ to obtain a watermarked signal $y(n)$. It is important for the transform domain method to ensure that the transform

domain samples, $Y(k)$, can take the inverse transform with appropriate forms. To demonstrate this requirement, let us consider the discrete Fourier transform (DFT) method. DFT should preserve the symmetric property of frequency domain samples within $[-\pi, \pi)$ in order to obtain real-valued samples of $y(n)$ after the inverse transform.

The spread spectrum (SS) method is one of the well-known transform domain methods. It spreads the message signal by a pseudorandom noise (PN) signal and adds it to the host signal [38]. This method can robustly detect the embedded messages from the watermarked signal against various signal processing. However, the principle of the spread spectrum method has a critical problem that is spreading of the spectrum reduces the sound quality. R.Namikawa and M. Unoki try to solve the sound quality problem by proposing a method that spectrally spreads a message by using linear prediction residue and embeds the spread spectrum of the message into the host signal [20]. Their method can improve the sound quality and maintain the robustness. However, it is a non-blind method since it requires linear prediction residue of the host signal.

Human auditory-based method

The information hiding is based on audio content and the human auditory system. For example, M.Unoki proposed information hiding using the human auditory system (HAS) by considering cochlear delay characteristics to embed watermarks. Cochlear delay refers to the non-uniform delays of wave propagation in the basilar membrane; thus, lower frequency components need more time to be perceived. [39]. R. Nishimura proposed another example that is a system to achieve information hiding in the audio signal by exploiting the properties of spatial masking and ambisonics [40]. Phase coding can also be considered as a human auditory based-method. Since the human auditory system is not sensitive to the absolute phase of the speech signal, this fact is utilized to construct the phase information hiding method. In this method, the absolute phase of the speech signal is replaced by the reference phase, which represents the secret information. All subsequent signals must change the absolute phase simultaneously to ensure the fixed relative phase between the signal. In the extraction process, the phase detection is done by using the synchronization mechanism [41, 42].

2.1.4 AIH applications

Information hiding technology has been raising attention since people were concerned about copyright and security. Along with advancing digital tools and technology, it is much easier to duplicate unauthorized digital products. In general, information hiding applications can be summarized in the aspects of data confidentiality, copyright protection, nonrepudiation, anti-fake, and data integrity [17, 26]. Details are as follows.

Data Confidentiality

Data confidentiality aims to prevent the transmitted data in a network from being captured by unauthorized users and to avoid intrusions by malicious attackers. Data confidentiality is a critical element of many aspects such as network security, political, military concerns, commercial, financial, and personal privacy matters. Sensitive data such as online banking transactions, personal privacy, and essential documents need to be protected while delivered through the network. Information hiding technology protects these data by not arousing any interest from attackers. Besides, some contents that are unwilling to be known by others can be hidden so that only the authorities can acquire the secret messages.

Copyright Protection

With the rapid growth of networking and digital technique, many digital services are provided through networks, and its consequence is that much important information sends through the network, and these data are easy to modify and duplicate. Thus this will enormously harm the service providers' profits, so that information hiding offers copyright protection to solve a problem as mentioned earlier.

Nonrepudiation

For nonrepudiation, information hiding aims to confirm the achievement on the Internet because neither side could deny what actions he or she made or the actions of the other party when utilizing activity on the network. The two parties of the transaction use information hiding technology to embed their feature marks into communicated messages. The feature marks can be considered secret messages, and the encrypted feature marks can be regarded as watermarks. This watermark must be permanent or not easy to remove. Thus the purpose of confirming the action can be achieved.

Anti-fake

Anti-fake marks can be appended through information hiding technology to confirm the authenticity of confidential files. Confidential files may be critical documents for the confidential work units or institutions, maps issued by a publishing house, various forms or contracts in business activities, and personal information.

Data Integrity

The primary purpose of data integrity verification is to confirm that the data had not been modified while being transmitted over the communication channel or in the stored procedure. To distinguish the modified media can be done using the fragile watermark method. The watermark should be fragile because it would be destroyed if any modification occurs.

2.2 Singular Spectrum Analysis

The Singular spectrum analysis, a time-series analysis method, was propose in 1986 by Broomhead and King [43]. SSA aims to decompose the original series into the sum of a small number of independent and interpretable components such as a slowly fluctuating trend, oscillatory components, and a structureless noise. Consider a signal in time series, $F = [f_0\ f_1 \ldots f_{N-1}]^{\mathrm{T}}$ where F is a signal of sufficient length N. The principal goal of SSA is to decompose the original signal into a sum of series where each component in this sum can be identified and used to interpret its characteristics. After finishing an interpretation or analysis of each component, these components are reconstructed into an original series.

The SSA technique consists of two complementary stages: decomposition and reconstruction, and each of them includes two separate steps. The embedding step and singular value decomposition step are in the decomposition stage; meanwhile, the grouping step and diagonal averaging step are in the reconstruction stage.Figure 2.3 illustrated the basic SSA. The following is a brief review of the methodology of the basic SSA technique.

2.2.1 Stage 1: Decomposition

First step: Embedding

Embedding can be considered as a mapping that transfers a one-dimensional time series signal into the multi-dimensional series. Let consider a time series

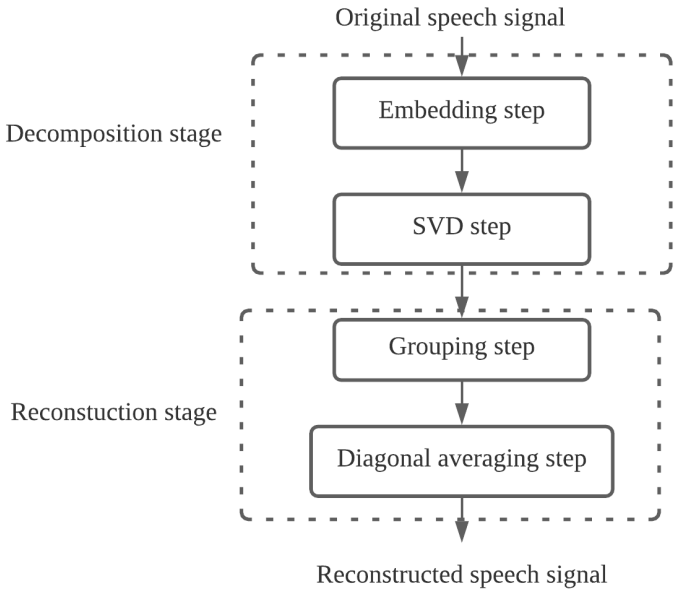


Figure 2.3: The basic SSA.

signal $F = [f_0\ f_1 \ldots f_{N-1}]^{\mathrm{T}}$ where f_i for $i = 0$ to $N-1$, where N is a total number of samples of the signal. A signal F is mapped to a trajectory matrix $\mathbf{X}$ of the size of $L \times K$, where L is the parameter of SSA, called a *window length*, and $2 \leq L \leq N$, and K is $N-L+1$, by the following relation.

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{K-1} \\ f_1 & f_2 & \cdots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \cdots & f_{N-1} \end{bmatrix}. \tag{2.3}$$

Vectors x_j are called *lagged vectors* and j^{th} is a column of matrix $\mathbf{X}$ i.e., $x_j = [f_j\ f_{j+1} \ldots f_{j+L-1}]^{\mathrm{T}}$. Thus $\mathbf{X} = [x_0\ x_1\ x_2 \ldots f_{K-1}]$. The result of this step is the trajectory matrix $\mathbf{X} = [x_0, \ldots, x_{K-1}]$. Note that the trajectory matrix $\mathbf{X}$ is a *Hankel matrix*, which can imply that all the elements throug the diagonal $i + j = \text{const}$ are equal. The procedure the embedding step is in time series analysis however future analysis depends on the aim of the investigation.

Second step: Singular value decomposition(SVD)

The SVD step makes the singular value decomposition of the trajectory matrix $\mathbf{X}$ and represents it as a sum of elementary matrices. Let consider the trajectory matrix $\mathbf{X}$ which SVD factorizes.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{p} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}, \tag{2.4}$$

where U_i and V_i are columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Note that U_i and V_i for $i = 0$ to $L-1$ are eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ respectively, and sorted in descending order. $\mathbf{\Sigma}$ is the diagonal matrix whose element are the square root of the eigenvalue of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, that can present as $\{\sqrt{\lambda_0}, \sqrt{\lambda_1}, ..., \sqrt{\lambda_p}\}$. we call λ_i for $i = 1$ to p a *"singular spectrum"*, and p is a number of positive eigenvalues greater than 0. Note that, in this work, we call a signal form by each $\sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ an *oscillatory component* of the signal F. Finally the trajectory matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{X_0} + \mathbf{X_1} + \mathbf{X_2} + ... + \mathbf{X_p}, \tag{2.5}$$

where $\mathbf{X_i} = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ and $p = \mathrm{argmax}_i(\lambda_i > 0)$

2.2.2 Stage 2: Reconstruction

First step: Grouping

The function of grouping step corresponds is to splitting the elementary matrices $\mathbf{X_i}$ into several groups, and then summing the matrices within each group. The set of indices of $\mathbf{X_i}$ obtained from the SVD step is partitioned into m disjoint subset I_l for $l = 1$ to m. Then the elements $X_1, X_2, ..., X_p$ are grouped in to m group, so that the trajectory matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{X_{I1}} + \mathbf{X_{I2}} + \mathbf{X_{I3}} + ... + \mathbf{X_I m}, \tag{2.6}$$

Second step: Diagonal averaging

. Diagonal averaging transfers each matrix $\mathbf{X_{Il}}$ into a time series. This step, also called Hankelization, map resultant matrix to a signal of length N by diagonal averaging (or Hankelization) [44]. The Hankelization of a matrix $\mathbf{Q}$ of size $L \times K$ to a signal $R = [r_0\ r_1\ ...\ r_{N-1}]^{\mathrm{T}}$, which is denoted by

$R = \mathcal{H}[\mathbf{Q}]$, is defined by the following equation.

$$r_k = \begin{cases} \frac{1}{k+1} \sum_{m=1}^{k+1} q^*_{m,k-m+2}, & \text{for} \quad 0 \leq k < L^* - 1, \\ \frac{1}{L^*} \sum_{m=1}^{L^*} q^*_{m,k-m+2}, & \text{for} \quad L^* - 1 \leq k < K^*, \\ \frac{1}{N-k} \sum_{m=k-K^*+2}^{N-K^*+1} q^*_{m,k-m+2}, & \text{for} \quad K^* \leq k < N, \end{cases} \tag{2.7}$$

where q_{ij} is an element at the row i and column j of the matrix $\mathbf{Q}$, $L^* = \min(L, K)$, $K^* = \max(L, K)$, $q^*_{ij} = q_{ij}$ when $L < K$, and $q^*_{ij} = q_{ji}$ when $L \geq K$.

2.3 Tampering Detection: state-of the art

When speech is transmitted through the network, it is possible that speech may be captured and tampered with by attackers. For example, changing speech content from the word "Yes" to "No" can mislead listeners, or conversion technique applied to speech can mislead about who is a speaker. Thus, tampering detection plays an essential role in many sensitive and vital cases as in digital forensic or severe problems for judgment in court. This paper proposes an information hiding method for tampering detection. Therefore, the definition of speech tampering and the scope of the work will be clarified in this section. In addition, the previously proposed methods used to solve the problem of tampering introduces in this section.

2.3.1 Tampering definition

The speech tampering can be classified into three groups [45]. The first is tampering with speech content, such as adding words, delete words, and replacing words. The second is tampering with speaker individuality by changing a characteristic of the speaker such as fundamental frequency (F0) changing, pitch shifting. Finally, the last is tampering with non-linguistic information, such as emotional information. Based on this classification, not all attacks should be viewed as tampering. Thus, two groups of attacks are malicious attacks (intentional modifications) and non-malicious attacks (unintentional modifications). The non-malicious attacks refer to normal signal operations such as re-sampling, re-quantization, and speech coding. The malicious attacks are those that change the speech content (e.g., replacement with unwatermarked segments), those that change the

speaker individuality or non-linguistic information of the speech signal (e.g., gradually speed changing, and significantly pitch shifting), and those that change the speaker environment (e.g., echo addition and noise addition to mislead about where speech is produced). The other operations, such as speech companding and compression, should not be considered as malicious attacks [46]. To correctly detect tampering, a watermark embedded into any host signal should be semi-fragile, i.e., the watermark is fragile to all malicious attacks but robust against non-malicious attacks or normal signal-processing operations [46].

There are other words whose meanings are close to the meaning of tampering. Those words are speech manipulation, speech modification, and speech spoofing. Let consider the spoofed speech carefully. There are four main spoofing types: Impersonation, replay, speech synthesis, and voice conversion [47]. The following gives a brief description of each type. Impersonation is related to human-altered voices, and it involves mainly mimicking prosodic or stylistic indications. Replay attacks are the presentation of speech samples taken from a genuine client in the form of continuous speech recordings or samples obtaining from the concatenation of shorter segments [48]. Voice conversion, a sub-domain of voice transformation, aims to convert one speaker's voice towards the target one. Voice conversion aims to synthesize a new speech signal that its features are close or similar to the target speaker. The last one is speech synthesis which requires large amounts of speaker-specific data to construct speech models that can produce a human-like voice. From the spoofing definition, note that there is an overlap between tampering and spoofing definitions. Voice conversion in spoofing attacks can be viewed as speaker individuality changing in tampering attacks. Spoofing speech by adding or replacing a synthesized voice can be viewed as changing speech content in a tampering attack. Since we want to prove the genuineness of speech signals, Thus, impersonation and replay attacks are out of this scope, and voice conversion and voice synthesis in spoofing categories can be included in tampering attacks. Therefore, if speech signals were manipulated, modified, or spoofed, it could be said that those speech signals were tampering. We will refer to those modifications, manipulations, and spoofing as tampering in speech signals. The tampering detection refers to the method to detect those tampering.

2.3.2 Tampering detection method

In the first attempt to provide security for speech signals, protecting signals from being stolen or modified is considered rather than tracing or detecting the modification. The method of scrambling and encryption is

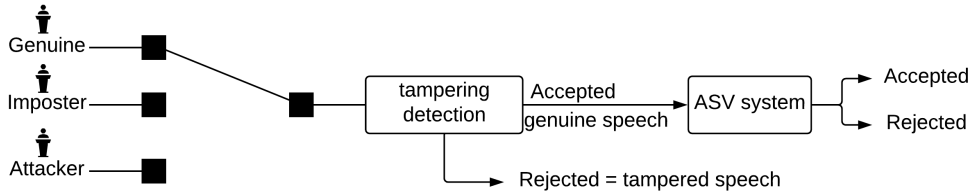


Figure 2.4: A tampering detection applied in ASV system.

adopted to protect. However, it can not be used for tampering detection.

In recent years, the spoofed speech, which can be considered as tampered speech, has gained more interest in the automatic speaker verification (ASV) system. An attacker attempts to bias the system outcome towards accepting a false identity claim in a tampering scenario. Thus, tampering detection, sometimes called countermeasure, gain attention in many systems, especially in automatic speaker verification system. Countermeasures or tampering detection have been developed to prevent attacks by deciding whether a particular trial is a genuine access attempt or a tampered one. Figure 2.4 illustrates a tampering detection applied in an automatic speaker verification system. Note that the tampering detection should decrease the fault accepted rate (FAR) in the event of tampering attacks while not increasing the fault rejected rate (FRR) in the case of genuine access attempts of the ASV system.

Figure 2.5 depicts a tampering detection system. The associated feature will be extracted from the speech signal and then be classified by a classifier. Results for the classified will be used to determine whether the speech is genuine or tampered speech. In this system, feature extraction and modeling are two key modules for detecting spoofed speech. There are several features to be used. For example, Chen et al. (2010) employed higher-order Mel-cepstral coefficients (MCEPs) to detect synthetic speech [49], De Leon et al. use F0 statistics for synthetic speech detection [50], and Wu et al. proposed a method to detect voice conversion by using Cosine normalized phase (cos-phase) and modified group delay phase (MGD-phase) features. Meanwhile, there are also several classifier models, for example, Gaussian mixture models (GMM)-based systems, Hidden Markov Model (HMM), and GMMs combined with a universal background model (UBM) to become GMM-UBM approach. However, this method requires large data for training and needs enormous computation time. Besides, there is no feature or modeling can apply for all kind of attacks.

Another approach to detecting tampering is to use the information hiding

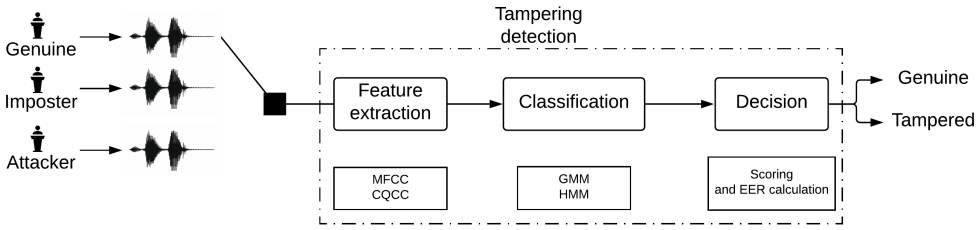


Figure 2.5: A tampering detection system.

method. Although the idea of information hiding appeared very early, it received much attention after 1996. Information hiding is a newly emerging information technology. The main driving is copyright protection. The information hiding for tampering detection gain more interest from a researcher in the past twenty years. The fragile watermarking techniques were used for tampering detection, for example, those used by Wu and Kuo [51]. For the purpose of tampering detection, information hiding scheme should be fragile only against malicious modifications and be robust to essential processing such as re-sampling and speech coding. Figure 2.6 depicts a tampering detection using information hiding.

The original speech, A, and embedded watermarks, w, are used on the sender side to produce the watermarked signal, A^*. The A^* on the receiver side can be received from the sender side and then the detection process blindly detects $\hat{w}$ from A^*. The tampering detection process verifies $\hat{w}$ with the shared w, to decide whether the signal is tampered with or not. Note that j is a function for making a decision.

Let us consider some information hiding method that use for tampering detection in speech and audio signals. A fragile speech watermarking scheme to detect malicious content was proposed by Wu, and Kuo [51]. The scheme used odd/even modulation with exponential scale quantization to detect tampering. Their system was able to distinguish tampering from resampling, white noise addition, G.711, and G.721 speech coding. However, it was not compatible well with CELP speech coders such as G.723.1. In 2012, M.Unoki and R.Miyauchi proposed information hiding for tampering detection based on cochlear delay characteristics [45]. Their method could detect tampering for content replacement, additive white noise, some malicious attack. However, their method was incorrect detect for normal signal operation such as G.711 speech coding. In 2014, Z.Liu and H.Wang proposed tampering detection in the content of speech signals [52]. Their method was based on Bessel–Fourier moments and the attacks such as insertion and deletion

of content were tested. However, in 2021, RCW.Phan reveals that this method has a weakness because their group can simulate the attacks and fault this system [16]. S.Wang et al. proposed a tampering detection scheme for speech signal using formant enhancement based method in 2015 [53]. This scheme modified line spectrum frequency for hiding information. Their method showed good ability to detect tampering, but there were also too fragile for some signal processing. The tampering detection using speech watermarking based on the source-filter model was proposed by S.Wang et al. in 2019 [54]. The speech was separated into the sound source and vocal tract information. Both were separately embedding with quantization index modulation and formant enhancing based technique. This scheme has a disadvantage to speech codec G.723.1, G.726, and G.729 affected by the QIM method. CO.Mawalim et al. proposed a watermarking method by modifying the least significant quantization bit in CELP codec [55]. This method was robust against CELP speech coding but low embedding capacity and fragile to some signal operation. The speech tampering detection using sparse representation and spectral manipulation-based information hiding was proposed in 2019 by S.Wang et al. [56]. Their method success in tampering detection and robust to speech coding.

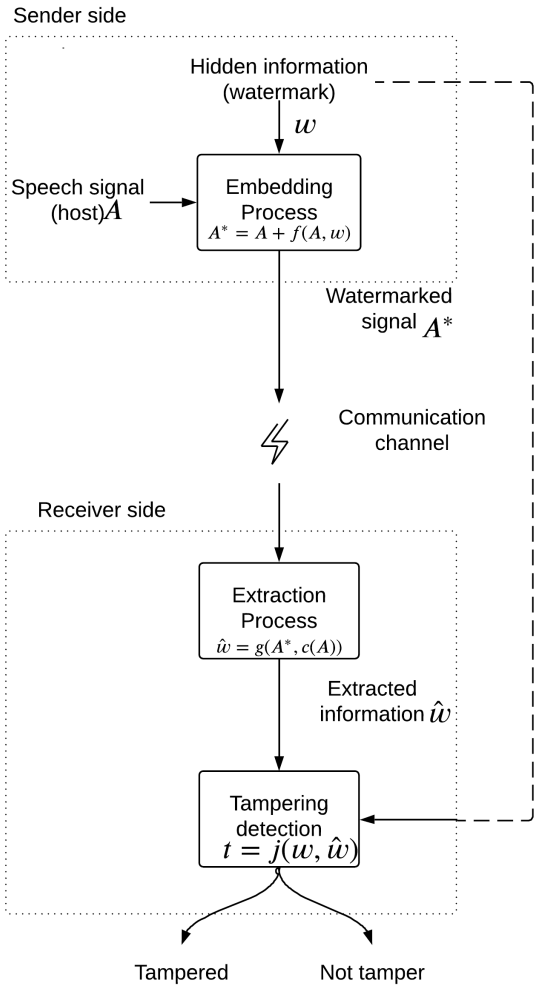


Figure 2.6: A tampering detection system using information hiding.

Chapter 3

SSA-based AIH core structure for tampering detection

3.1 SSA-based AIH for tampering detection and its issues

SSA-based information hiding was proposed by Karnjana et al. in 2014 [57]. SSA-based information hiding adopts the analysis of the SVD-based technique as its core structure and considers the relationship between singular value and human perception [58]. The information hiding scheme used basic SSA to analyze host signals and extract the singular spectra, and a watermark was hidden into a host signal by selecting a part of the singular spectrum of the host signal and then modifying the selected part concerning the watermark bit. Our studies discovered that the SSA-based information hiding scheme could be made robust, fragile, or semi-fragile depending on which part of the singular spectrum we selected to modify. The modification affects the sound quality of the watermarked signal and the robustness of the information hiding scheme. Therefore, a part of the singular spectrum to be modified must be determined appropriately to balance inaudibility and robustness. The parameter used to identify the specific interval of singular spectra is called as *embedding parameter*, or parameter as short.

There are many empirical experiments to finding these parameters. Figure 3.1 shows the example of a singular spectrum of one frame speech signal. The window with the dashed line displays the part on the singular spectrum to be modified. This window is moved along the x-axis to check whether the sound quality and robustness change when the window moved. In other words, to check the balance of inaudibility and robustness when embedding position change. From this experiment, we found that if the watermark

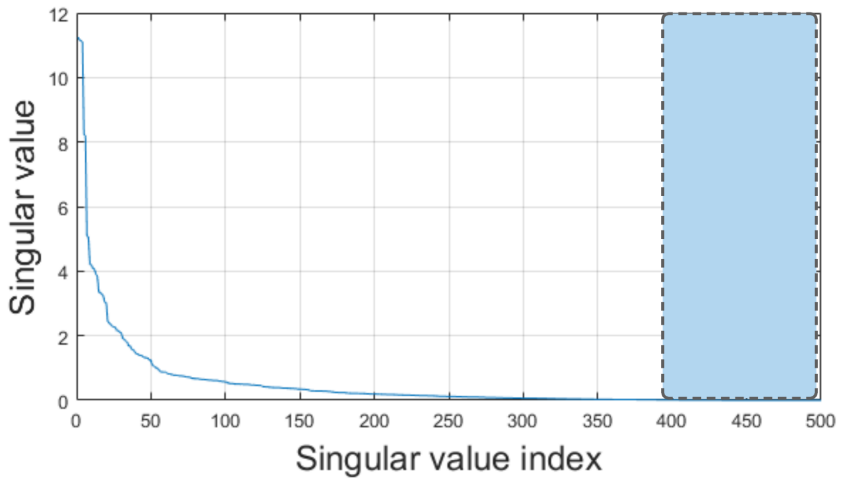


Figure 3.1: Singular spectrum of one frame.

is embedded in a low index of the singular value, it is easy to extract the watermark in the extraction process. However, the sound quality is degraded significantly. Simultaneously, if the watermark is embedded in a high index of the singular value, the sound quality is excellent, but it is difficult to extract the watermark in the extraction. Note that the watermark is more fragile if we hide it in a high singular value index. Therefore, the interval of the singular spectrum to be modified must be determined appropriately not only for balancing inaudibility and robustness but also to satisfy the semi-fragility. Let us consider the characteristics of the singular spectrum of the different speech segments. Figure 3.2 showed the singular spectrum of the four different speech segments. It can be seen that the singular spectrum is different for each speech segment. Consequently, the embedding parameter of each frame will also be different. Therefore, parameter estimation is dependent on speech segment.

3.2 Philosophy of this work

In this work, we construct the core structure of the watermarking scheme for detecting tampering. We get the motivation from Karnjana et al. research that proposed differential evolution (DE) optimization to determine such a suitable window for balancing inaudibility and robustness [59]. Therefore, we hypothesize that DE should be able to find a parameter to satisfy the

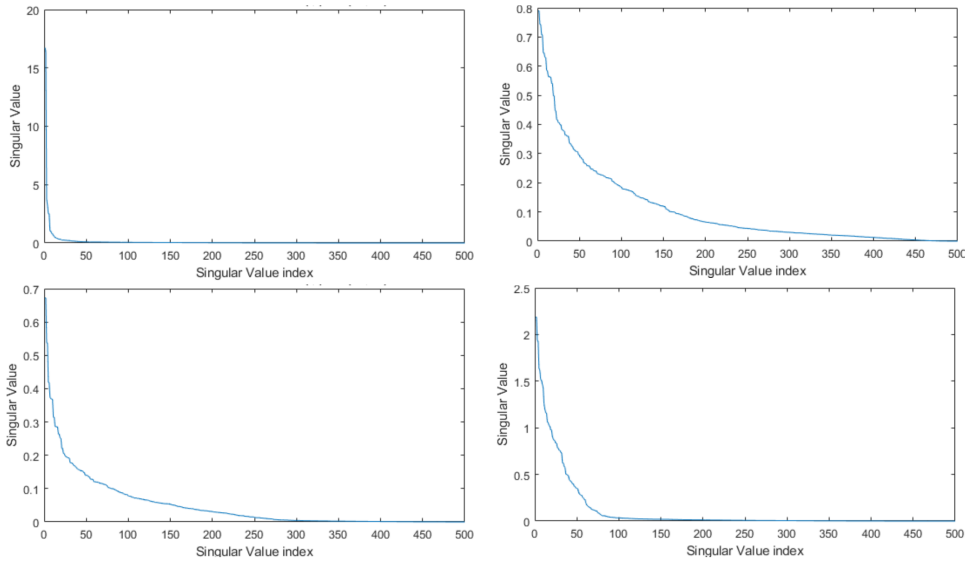


Figure 3.2: Singular spectrum of the four different speech segments.

semi-fragility. However, DE requires huge computation time and cannot be practically used in real-time or near real-time applications. Therefore, We deploy a neural network to estimate the deterministic relationship of the input speech signal and parameters that are used to identify the suitable part of the singular spectrum of the speech signal. We propose a unique convolutional neural network (CNN)-based parameter estimation method to estimate the parameter used for the information hiding scheme. Because the performance of a neural network depends considerably on the dataset used to train the neural network, the important ingredient of this work is the framework we use to create a good dataset. As mentioned earlier, DE has proved its excellence in the trade-off between inaudibility and robustness. We presume that it can effectively be used as a foundation for generating a training dataset.

3.3 The core structure of SSA-based AIH for tampering detection

The proposed information scheme is based on the framework of singular spectrum analysis (SSA) and consists of two main processes, an embedding process, and an extraction process, as depicted in Fig. 3.3. It is a blind

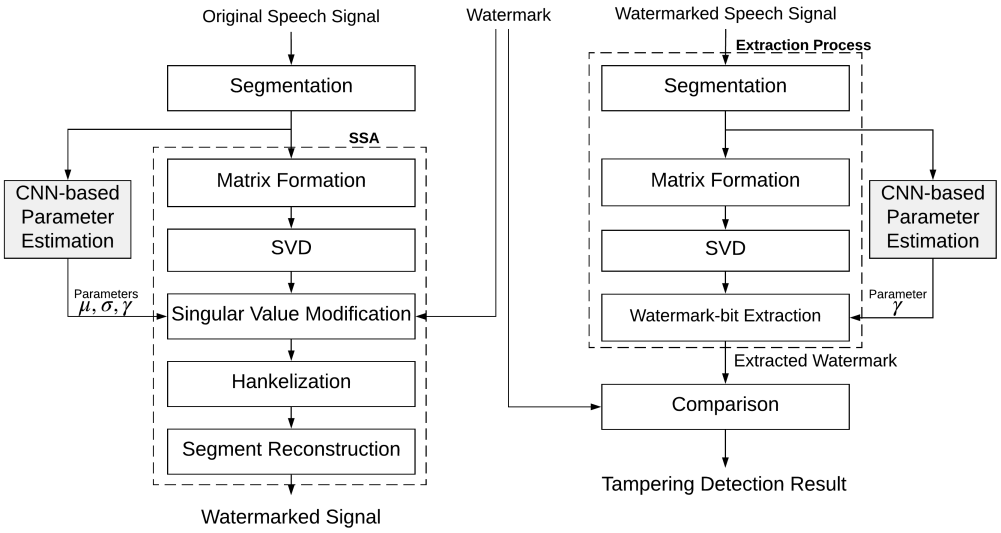


Figure 3.3: Proposed framework: embedding process (left) and extraction process with tampering detection (right).

scheme where its extraction process can extract hidden information from only a watermarked signal i.e., no need host signal on the extraction process. Also, the extraction process is parameter-free because all parameters can be estimated from the watermarked signal by using a CNN-based algorithm. In other words, no need to share parameters between the embedding and extraction process.

This section summarily gives details on these two processes and how to use them for detecting tampering.

3.3.1 Embedding Process

The embedding process provides a watermarked signal by taking a host signal and a watermark as its inputs, and one frame of host signal will be embedded with one watermark bit. The embedding process consists of six steps, as shown in Fig. 3.3 (left), which are described as follows.

1. *Segmentation.* A host signal is segmented into frames of length N, where M is equal to the total number of watermark bits. Note that frame type is non-overlapping frame.Let F indicates a segment of length N, i.e., $F = [f_0\ f_1 \dots f_{N-1}]^{\mathrm{T}}$, where samples of a segment is f_i for $i = 0$ to $N-1$.

2. *Matrix Formation.* A segment F is mapped to a matrix $\mathbf{X}$ with the following equation.

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix}, \tag{3.1}$$

where L, which is named a "window length" of the matrix transformation. The lowest value of L is 2 and its value is not greater than N. The size of $\mathbf{X}$ is $L \times K$, where $K = N - L + 1$.

3. *Singular Value Decomposition (SVD).* We factorize the real matrix $\mathbf{X}$ by using SVD, i.e.,

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \tag{3.2}$$

where the columns of $\mathbf{U}$ and those of $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix that whose elements are the square roots of the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

Denote $\sqrt{\lambda_i}$ for $i = 1$ to q indicate the elements of $\mathbf{\Sigma}$ in descending order and $\sqrt{\lambda_q}$ is the smallest eigen value which is non-zero. The $\sqrt{\lambda_i}$ is called a "singular value" and the $\{\sqrt{\lambda_0}, \sqrt{\lambda_1}, ..., \sqrt{\lambda_q}\}$ is called a "singular spectrum."

4. *Singular Value Modification.* The singular spectrum is modified regarding the watermark bit to be embedded and requires the properties of the watermarking scheme. Our previous work shows that modifying high-order singular values is less distorts the host signal but is sensitive to noise or attacks. Meanwhile, modifying low-order singular values can improve robustness but causes sound quality to be poor [59, 60]. Thus, there is a trade-off between the sound quality of the watermarked signal and the scheme's robustness. In this work, we aim mainly on semi-fragility property. Therefore, we introduce the embedding rule as follows.

A whole singular spectrum is defined by $\{\sqrt{\lambda_0}, \sqrt{\lambda_1}, ..., \sqrt{\lambda_q}\}$. A specific part of this singular spectrum, presented as $\{\sqrt{\lambda_p}, \sqrt{\lambda_{p+1}}, ..., \sqrt{\lambda_q}\}$, is modified regarding to the embedded-watermark bit w. The modification rule is as follows.

$$\sqrt{\lambda_i^*} = \begin{cases} \sqrt{\lambda_i} + \alpha_i \cdot (\sqrt{\lambda_p} - \sqrt{\lambda_i}), & \text{if } w = 1, \\ \sqrt{\lambda_i} \quad \text{(i.e., unchanged)}, & \text{if } w = 0, \end{cases} \tag{3.3}$$

where $\sqrt{\lambda_i^*}$ is the singular values was modified for $i = p$ to q. The $\sqrt{\lambda_p}$ is the largest singular value that is less than $\gamma \cdot \sqrt{\lambda_0}$, and α_i, which is called the "embedding strength," is normally distributed over the interval $[p, q]$ and the embedding strength has a maximum value of 1. Note that α_i is a positive real value that is less than 1. Specifically, α_i for $i = p$ to q is determined by

$$\alpha_i = e^{-(i-\mu)^2/2\sigma^2}, \tag{3.4}$$

where μ and σ^2 are the mean and the variance of the Gaussian distribution, respectively.

Hence, the embedding rule requires three parameters, which are γ, μ, and σ. From the empirical experiment, we have shown that by appropriately adjusting these parameters concerning the host signal, we can deliver a good balance between watermarked signal 's sound quality and robustness of the scheme [60].

The core structure of SSA-based AIH with CNN-based parameter estimation is shown in Fig. 3.3. The left-hand side shows, these parameters are provided by the CNN-based parameter estimation, which is to be discussed in detail in the following section. Figure 3.4 showed an example of the part $\{\sqrt{\lambda_p}, \sqrt{\lambda_{p+1}}, ..., \sqrt{\lambda_q}\}$ of a singular spectrum.

5. *Hankelization.* Let $\mathbf{\Sigma}^*$ denote a diagonal matrix defined by

$$\mathbf{\Sigma}^* = \begin{bmatrix} \sqrt{\lambda_0} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \sqrt{\lambda_{p-1}} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sqrt{\lambda_p^*} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sqrt{\lambda_q^*} \end{bmatrix}. \tag{3.5}$$

A watermarked matrix $\mathbf{X}^*$ is the matrix into which the watermark bit is embedded can be computed from the product $\mathbf{U}\mathbf{\Sigma}^*\mathbf{V}^{\mathrm{T}}$. Then, a hankelization is performed on the matrix $\mathbf{X}^*$ to obtain the signal F^*, as the watermarked segment. A hankelization is the average of the anti-diagonal $i+j=k+1$, where i is the row index and j is the column index of matrix of an element of $\mathbf{X}^*$, and k (for $k=0$ to $N-1$) is the index of element F^*.

6. *Segment Reconstruction.* We obtain the watermarked segment from the previous step, and then on final step, all watermark segments is sequentially concatenated to produce the watermarked signal.

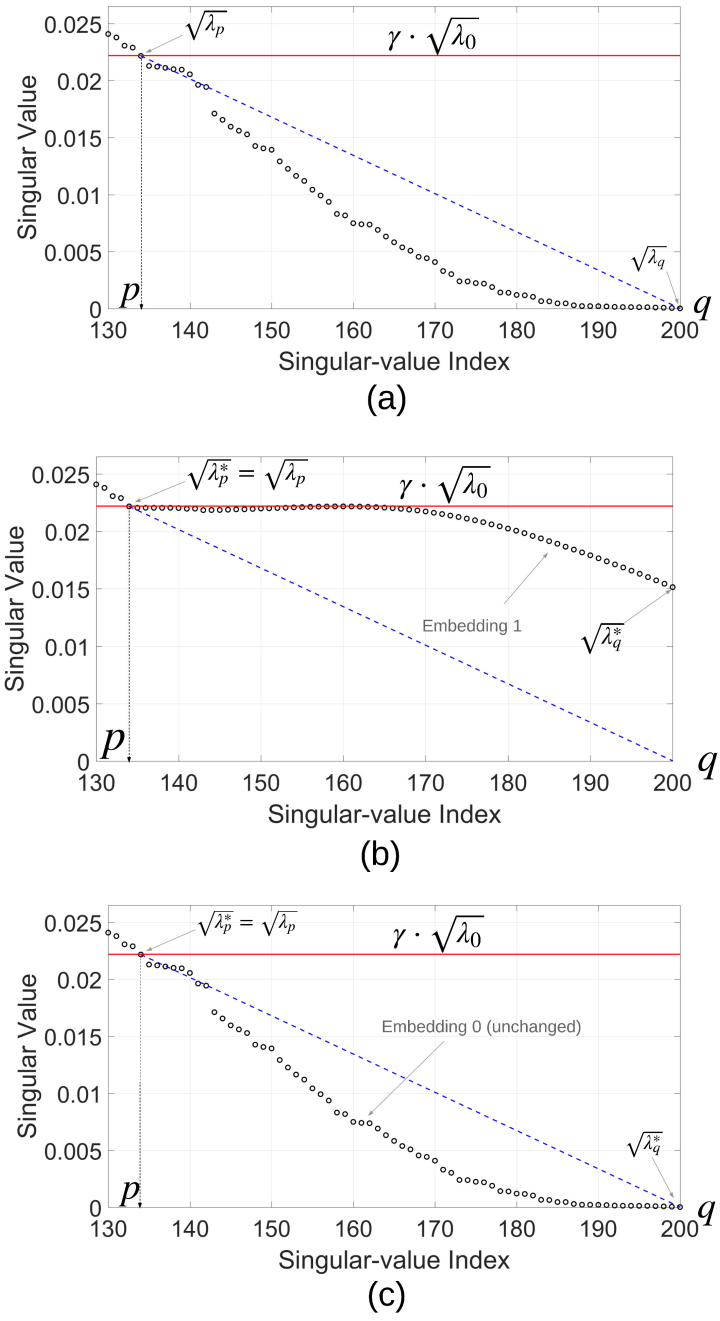


Figure 3.4: Example of the part of a singular spectrum $\{\sqrt{\lambda_p}, \sqrt{\lambda_{p+1}}, ..., \sqrt{\lambda_q}\}$: (a) selected part of singular spectrum without embedding, (b) watermark bit 1 is embedded, and (c) watermark bit 0 is embedded. The red line shows the threshold level $\gamma \cdot \sqrt{\lambda_0}$, and the blue dashed line connects from $\sqrt{\lambda_p}$ to $\sqrt{\lambda_q}$.

3.3.2 Extraction Process

The extraction process uses a watermarked signal to extract an embedded watermark and delivers an extracted watermark signal as its output. There are four steps in the extraction process, as shown in the dashed box of Fig. 3.3 (right). Three first steps are the same as those describes the embedding process, which are *segmentation*, *matrix formation*, and *singular value decomposition*. The fourth step is *watermark-bit extraction*. Embedded watermark bits are extracted in this step by decoding singular spectra. How the spectra are decoded is depends on how they are modified in the embedding process. To explain the idea of decoding, let us consider the two singular spectra in Fig. 3.5. This figure displays two extracted singular spectra of one watermarked frame when embedded with different watermark bits: $\{\sqrt{\lambda^*_{0^0}}, ..., \sqrt{\lambda^*_{p^0}}, ..., \sqrt{\lambda^*_{q^0}}\}$ and $\{\sqrt{\lambda^*_{0^1}}, ..., \sqrt{\lambda^*_{p^1}}, ..., \sqrt{\lambda^*_{q^1}}\}$. The superscripts of the indices of singular values, 0 and 1 , indicate the embedded watermark bits. It can be seen that most singular values (circles) under the red line are above the blue dashed line connecting $\sqrt{\lambda_p}$ and $\sqrt{\lambda_q}$, when the watermark bit 1 is embedded, but most of the singular values (asterisks) under the red line are below the blue dashed line when the watermark bit 0 is embedded. Therefore, we can use the following state to determine the hidden watermark bit $\hat{w}$.

$$\hat{w} = \begin{cases} 0, & \text{if } \sum_{i=p}^{q} \left(\sqrt{\lambda^*_i} - l(i)\right) < 0, \\ 1, & \text{if } \sum_{i=p}^{q} \left(\sqrt{\lambda^*_i} - l(i)\right) \geq 0, \end{cases} \tag{3.6}$$

where $l(i)$ is the corresponding values on the blue dashed line, which is defined by

$$l(i) = \left(\frac{\sqrt{\lambda^*_p} - \sqrt{\lambda^*_q}}{p-q}\right) \cdot (i-q) + \sqrt{\lambda^*_q}. \tag{3.7}$$

The output of the fourth step is the extracted watermark bit $\hat{w}(j)$ for $j = 1$ to M.

3.3.3 Tampering Detection

The concept to check whether watermarked signals have been tampered with or not, we compare extracted-watermark bits $\hat{w}(j)$ with embedded-watermark bits $w(j)$ for $j = 1$ to M to see its matching. For the purpose

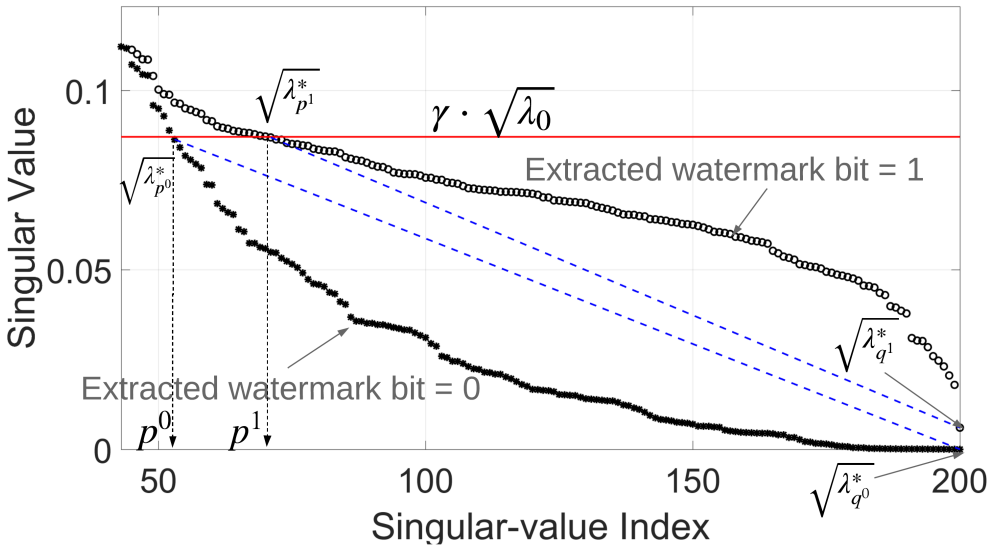


Figure 3.5: Decoding hidden watermark bit: if most of singular values (circle) that are under threshold level $\gamma \cdot \sqrt{\lambda_0}$ are above blue dashed line, extracted watermark bit is 1, but if most of singular values (asterisks) that are below threshold level $\gamma \cdot \sqrt{\lambda_0}$ are under blue dashed line, extracted watermark bit is 0.

of tampering detection, the embedded-watermark bits $w(j)$ are assumed to be known by the owner or an authorized person. Theoretically, when the tamper occurs, watermark bits embedded into the tampering location should be destroyed. Thus, tampering and spoofing could be detected by mismatches between $\hat{w}(j)$ and $w(j)$. Because we embed one watermark bit into one frame of the host signal, then each mismatch position can be used to indicate the corresponding tampered frame. Therefore, tamper position can be identified.

3.4 CNN-based Parameter Estimation

As mentioned earlier, we previously proposed a watermarking scheme in which an evolutionary-based optimization algorithm, differential evolution (DE), was deployed to find input-dependent parameters used in the embedding process of the scheme [59]. In that work, the method of determining input-dependent parameters is called as "parameter estimation." We discovered that our DE-based parameter estimation could produce parameters that result in a good balance between the robustness and inaudibility of that proposed scheme [59]. However, the computing time of DE-based parameter

estimation is significant high [61, 62, 63]. To reduce this computational time, we consequently proposed another approach based on a convolutional neural network (CNN) [64]. Using this CNN-based parameter estimation, we hugely reduced the computational time by approximately $10,000$ times [64]. Although we succeeded in reducing the computational time, we had to sacrifice robustness in this previous work. We hypothesize that if we use the high-quality dataset for CNN training, it can yield this problem. Accordingly, in this work, we develop the CNN-based parameter estimation by improving the quality of the CNN training dataset. In this section, we explain how we provide a high-quality dataset and an enhanced CNN-based approach.

There are two crucial steps to implementing the improved CNN-based parameter estimation: CNN training and generating a high-quality dataset. The details of these two steps are detailed in the following subsections.

3.4.1 Training CNN

A CNN is a feed forward neural network which had supervised learning and unsupervised learning. In this work, we implement a supervised learning scheme that CNN is trained by a training dataset consisting of various input and target pairs. These pairs of input and target are used to find a deterministic function that maps an input to obtain an output, and the trained CNN achieves this function [65].

Simply saying, the CNN is used to find the vital embedding parameters γ, μ, and σ for each speech segment. We choose the CNN in this work because we know that there is a relationship between singular values and speech signal frequencies [61, 63]. For example, high-order singular values are associated with a high-frequency band; meanwhile, low-order singular values are associated with a low-frequency band. Thus, we hypothesize that a CNN trained with inputs represented in both time and frequency domains should perform better than the one with either a CNN trained with only time-domain input or with only frequency-domain input. Thus, the spectrograms of the speech segments are chosen as the inputs in the training dataset. Considering a spectrogram is two-dimensional (2D), and the CNN can extract patterns in 2D data more efficiently than other neural networks. We, therefore, designed our novel parameter estimation based on the CNN.

As mentioned in the earlier section, there are three parameters, γ, μ, and σ, to be optimized. Two of these parameters, μ and σ, associate with the embedding strength α_i. Thus, they provide the robustness of the proposed scheme. The parameter γ directly determines the number of modified singular values. Simply saying, it contributes more to the sound quality aspect of the proposed scheme. These two groups are different in

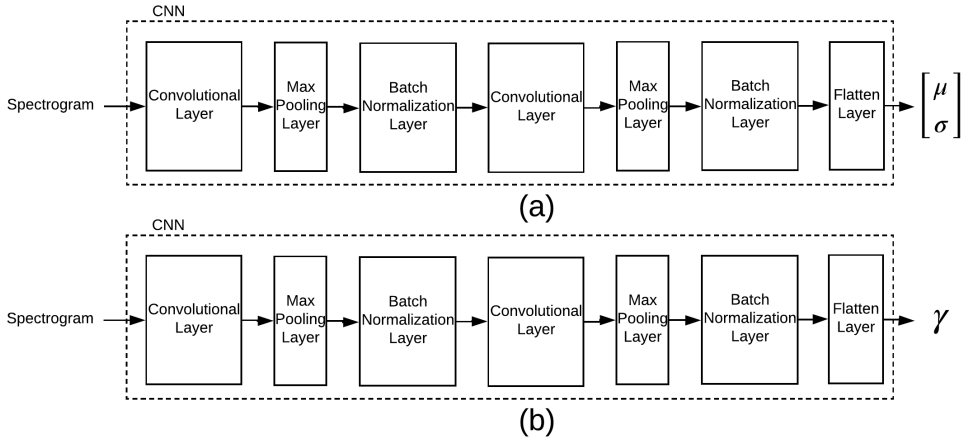


Figure 3.6: Structure of two CNNs: (a) CNN used to estimate embedded strength parameters and (b) CNN used to estimate parameter of γ.

terms of value, i.e., γ has a small value than another. Consequently, two CNNs were implemented, one for μ and σ and the other for γ. A spectrogram of size 13×67 is the input of both CNNs. Our CNNs are composed of two convolutional layers, two pooling layers, and two normalization layers. A first convolution layer convolutes an input spectrogram with 128 kernels of size 3×3 and a stride of size 2×2, and the other convolutes with 64 kernels of size 3×3. The activation function for this CNN is a rectified linear unit (ReLU) function. A kernel size of 2×2 is deployed for all pooling layers. A flattened output is combined with a fully connected layer with 256 units. The given outputs of the first CNN and the second CNN are $[\mu\ \sigma]^{\mathrm{T}}$ and the parameter γ, respectively. The structure of both CNNs is presented in Fig. 3.6.

3.4.2 Generating High-Quality Dataset

We discovered that DE proved its effectiveness in finding the excellent parameters in our previous work [59], we, therefore, deploy it to create a dataset for supervising our CNNs. We gives a definition of a high-quality dataset in this proposed method as a dataset in which a good sample of input-output pairs used for CNN supervising so that the CNN can map from the input and particular output with high-precision estimation. DE functions as follows.

Let $\mathbf{x}$ be a D-dimensional vector that we want to find according to a cost function $C(\mathbf{x})$, i.e., we are searching for $\mathbf{x}$ such that $C(\mathbf{x})$ is minimized. The

algorithm of DE consists of four steps: initialization, mutation, crossover, and selection [66].

First, we initialize vectors $\mathbf{x}_{i,G}$, for $i=1$ to NP, where NP is a size of the population in the generation G. For the initialization step, $G=1$.

Second, each $\mathbf{x}_{i,G}$ is mutated to a vector $\mathbf{v}_{i,G+1}$ by $\mathbf{v}_{i,G+1} = \mathbf{x}_{r_1,G} + F \cdot (\mathbf{x}_{r_2,G} - \mathbf{x}_{r_3,G})$, where i, r_1, r_2, and r_3 are distinct and random from $\{1, 2, .., NP\}$. The predefined constant F is in the interval $[0, 2]$ and used to determines the convergence rate of DE.

Third, each pair of $\mathbf{x}_{i,G}$ and $\mathbf{v}_{i,G+1}$ is used to generate another vector $\mathbf{u}_{i,G+1}$ by using the following formula. Given that

$$\mathbf{u}_{i,G+1} = \begin{bmatrix} u_{1i,G+1} & u_{2i,G+1} & ... & u_{Di,G+1} \end{bmatrix}^{\mathrm{T}}, \tag{3.8}$$

$$\mathbf{u}_{ji,G+1} = \begin{cases} \mathbf{v}_{ji,G+1}, & \text{if } \Xi(j) \leqslant CR \text{ or } j = \upsilon, \\ \mathbf{x}_{ji,G}, & \text{otherwise,} \end{cases} \tag{3.9}$$

$\Xi(j)$ is a random real number in the interval $[0, 1]$, CR is a predefined constant in $[0, 1]$, and υ is random from $\{1, 2, ..., D\}$.

In the last step, we compare $C(\mathbf{x}_{i,G})$ with $C(\mathbf{u}_{i,G+1})$. If $C(\mathbf{x}_{i,G}) < C(\mathbf{u}_{i,G+1})$, $\mathbf{x}_{i,G+1} = \mathbf{x}_{i,G}$; otherwise, $\mathbf{x}_{i,G+1} = \mathbf{u}_{i,G+1}$. After obtaining all members of the generation $G+1$, then we iteratively repeat those three steps, mutation, the crossover, and the selection step, until our specific condition is satisfied. Then, the DE algorithm gives $\mathbf{x}_i$, which yields the lowest cost in the final generation as the answer.

A DE optimizer used for creating the dataset is shown in Fig. 3.7. Note that our DE optimizer is included with a few compression algorithms and coding algorithms because we want to ensure that our proposed scheme is robust against these operations. The extraction processes in Fig. 3.7 are a bit different from the extraction process described for the SSA-based AIH core structure. The difference is all extraction processes in the DE optimizer know the parameter γ used in the embedding process, while the extraction process in the SSA-based AIH core structure is completely blind.

We defines the cost function C developed in this work as follows.

$$\begin{aligned} C = {} & \beta_1 \mathrm{BER}_{\mathrm{NA}} + \beta_2 \mathrm{BER}_{\mathrm{MP3}} + \beta_3 \mathrm{BER}_{\mathrm{MP4}} \\ & + \beta_4 \mathrm{BER}_{\mathrm{G711}} + \beta_5 \mathrm{BER}_{\mathrm{G726}}, \end{aligned} \tag{3.10}$$

where β_i for $i=1$ to 5 are constants and $\sum_{\forall_i} \beta_i = 1$, and BER is the bit-error rate. The BER can be used to express the extraction precision and is defined as

$$\mathrm{BER} = \tfrac{1}{M} \sum_{j=1}^{M} w(j) \oplus \hat{w}(j), \tag{3.11}$$

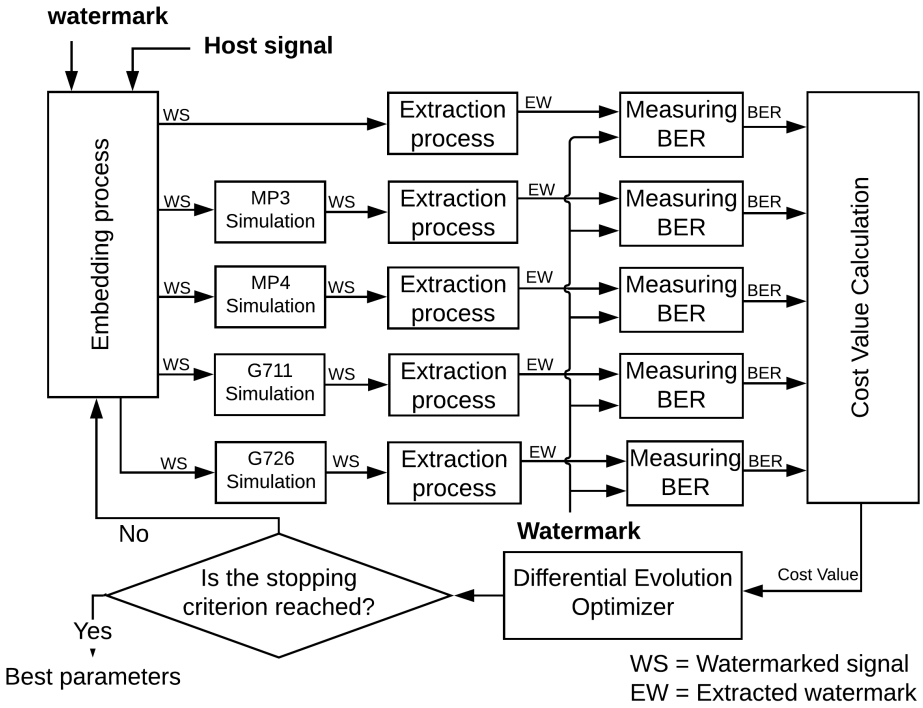


Figure 3.7: DE optimizer used to create dataset

where $w(j)$ and $\hat{w}(j)$ are the embedded-watermark bits and the extracted-watermark bits, respectively, and the symbol $\oplus$ presents a bitwise XOR operator. Hence, the terms $\mathrm{BER_{NA}}$, $\mathrm{BER_{MP3}}$, $\mathrm{BER_{MP4}}$, $\mathrm{BER_{G711}}$, and $\mathrm{BER_{G726}}$ denote the average BER values when there is no attack, when MP3 operation is performed, when MP4 operation is performed, when G.711 speech coding is performed, and when G.726 speech coding is performed on watermarked signals, respectively.

Note that, although our selected cost function is a function of only BERs, we can set the upper bound of the parameter γ in the DE algorithm to control the sound quality of watermarked signals. Issues regarding the cost function will be addressed in more detail after we have shown our evaluation results. The framework used to create the training dataset is shown in Fig. 3.8.

In the embedding process, the host speech signal is segmented into non-overlapping frames. Each frame will be embedded with one watermark bit. Thus, the frame length reflects the embedding capacity, and the number of frames, M, is equal to the number of the watermark bits to be embedded. Then the trajectory matrix $\mathbf{F}$ which represents each frame F is created. Each trajectory matrix $\mathbf{F}$ is performed with Singular value decomposition (SVD) to obtain each frame's singular spectra. The singular spectra are then modified

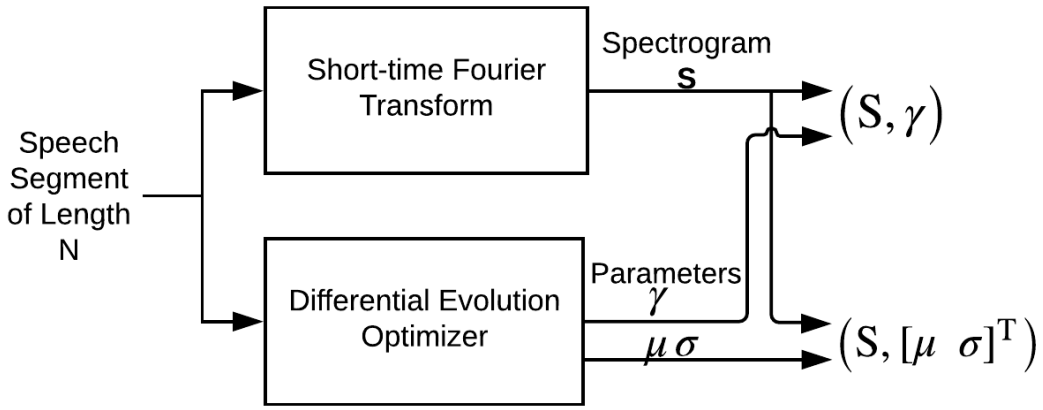


Figure 3.8: Framework for creating training dataset.

to hide the watermark bit (0 or 1), and the part of the singular spectra to be modified depends on the necessity of the information hiding application. The modified trajectory matrix $\mathbf{Y}$ is constructed by SVD reversion and then hankelized. The hankelization of a modified trajectory matrix $\mathbf{Y}$ yields a signal G, where G is a frame of the watermarked signal. The frames are stacked to reconstruct the watermarked signal.

In the extraction process, the watermarked signal is segmented into non-overlapping frames, and the trajectory matrix for each frame is constructed in the same way as was done in the embedding process. Then SVD is performed on the trajectory matrix to obtain the singular spectra. The singular spectra of the signal of each frame are typically convex; however, the watermark bit embedded into a part of the singular spectrum of a host frame results in a concave on the part of the singular spectrum of the reconstructed, watermarked frame. The concave and convex caused by the embedding process can be utilized to extract the watermark bit from each frame.

Chapter 4

Evaluation and Results

This chapter provides dataset, evaluation condition, and results. We evaluated the proposed scheme concerning four aspects: the sound quality of watermarked signals, semi-fragility, tampering detection ability, and computational time. The criteria to evaluate the proposed method has complied with the requirement of information hiding committees on the paper named "Information Hiding and Its Criteria for Evaluation" [67]. The performance from evaluation results is compared with our previously proposed schemes [46, 60]. Moreover, the proposed scheme also compare with three other conventional methods: a method based on time-domain information hiding where embedding information into the least significant bit (LSB) [68], a cochlear-delay-based (CD-based) method proposed by M.Unoki et al. [45], and a formant-enhancement based (FE-based) method proposed by S.Wang et al. [69].

4.1 Dataset and Conditions

Twelve speech stimuli from the ATR database B set (Japanese sentences uttered by six males and six females) were used as the host signals to evaluate the SSA-based AIH core structure for tampering detection [70]. We choose this dataset because we want to make a fair comparison between our previous methods and this proposed core structure. All signals are one channel with a sampling rate of 16 kHz, 16-bit quantization. The frame size was 25 ms or 40 frame in one second. Thus, there were 400 samples for one frame. In other words, our embedding capacity was 40 bps. Each signal was embedded with one hundred and twenty bits in total, and the embedding duration of each signal was three seconds. We used 200 diverse frames from each host signal to prepare the dataset for training the CNNs. Therefore, there were

2, 400 segments in our training dataset.

The hyperparameters for the DE algorithm in our simulation were set as follows. The population size in each generation (NP) was 30, as suggested by Storn *et al.* [66]. A maximum number of generations [max(G)] was 30. The upper bounds of the parameters γ, μ, and σ were set as 0.0085, 220, and 150, respectively. The lower bounds were set as 0.001, 80, and 0, respectively. The two constants F and CR were set as 0.5 and 0.9, respectively, as suggested by Storn *et al* [66]. The weights β_i in the cost function were set as follows. $\beta_1 = \frac{1}{3}$, $\beta_2 = \frac{4}{21}$, $\beta_3 = \frac{4}{21}$, $\beta_4 = \frac{4}{21}$, and $\beta_5 = \frac{2}{21}$.

In addition to the frame size N, which is 400, our proposed scheme requires another hyperparameter, i.e., the window length of the matrix formation (L). We set the window length L to one-half of the frame size in all simulations, which was 200.

4.2 Sound Quality Evaluation

We used three objective measurements: the log-spectral distance (LSD), the perceptual evaluation of speech quality (PESQ), and the signal-to-distortion ratio (SDR) to evaluate the speech quality of watermarked signals. The LSD, expressed in dB, is a distance between two spectra: the spectra of the original signal and the watermarked signal. The LSD is defined by

$$\text{LSD} = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[10 \log \frac{P(\omega)}{P(\hat{\omega})} \right]^2 d\omega}, \tag{4.1}$$

where $P(\omega)$ is the spectra of the host signal,and and $P(\hat{\omega})$ is the spectra of watermarking signal.

The PESQ measures the distortion of a watermarked signal compared with the host signal [71]. The PESQ score ranges from [−0.5-4.5], where (−0.5) indicate as very annoying and (4.5) indicate imperceptibly. The PESQ software used in this experiment is recommended by the International Telecommunication Union (ITU) for measurement in our experiment [72].

The SDR, expressed in dB, is the power ratio between the host signal and the distortion in watermarked signal, which is defined by

$$\text{SDR} = 10 \log \frac{\sum_n [A(n)]^2}{\sum_n \left[A(n) - A^*(n) \right]^2}, \tag{4.2}$$

where $A(n)$ is the amplitudes of the host signal, and $A^*(n)$ is the amplitudes of watermarked signals.

Table 4.1: Sound-quality evaluations: proposed scheme vs. other methods.

	PESQ score	**LSD** (dB)	**SDR** (dB)
LSB-based method [68]	4.49	0.19	65.35
CD-based method [45]	∼3.1-4.3	∼0.6-0.8	-
FE-based method [69]	∼3.9	∼0.4	-
SSA-based method (fixed parameters) [60]	3.64	0.69	30.96
SSA-based method (with ad-hoc parameters) [46]	3.70	0.65	31.58
Proposed method	4.05	0.45	35.51

The criteria for good sound quality in this work were set as follows. The LSD should not greater than 1 dB, a PESQ score of 3.0 was set as sufficient quality, and the SDR should be greater than 30 dB, as recommended in S.Wang's experiment. [53].

Table 4.1 showed the results of the sound quality evaluation. All information hiding methods satisfied the criteria for good sound quality. Note that besides the LSB-based method, our proposed one achieves a better performance than the others. Note that the proposed scheme was improved considerably in terms of sound quality compared to the previously proposed one.

4.3 Semi-fragility Evaluation

An information hiding scheme should be robust against non-malicious speech processing, for example, compression and speech coding, and fragile to malicious attacks, e.g., pitch shifting and band-pass filtering o detect tampering. The scheme's robustness can be indicated by the bit-error-rate (BER), as defined in (5.6). In this work, we set a threshold of a BER to be 10% as a robustness indication. A scheme is considered to be a robust scheme if its BER is less than 10%. If its BER is higher than 20%, the speech signal is considered as tampered speech. In the case of BER is between 10% and 20%, the speech signal is probably tampered with at a low degree or unintentionally modified [46].

The semi-fragility of the proposed scheme was evaluated by performing ten signal processing operations on the watermarked signals as follows. Four non-malicious operations: G.711 speech coding, G.726 coding, MP3 operation with 128 kbps, and MP4 operation with 96 kbps. Six possible malicious operations: band-pass filtering (BPF) with 100-6000 Hz and -12 dB/octave, Adding white Gaussian-noise (AWGN) with 15-dB and 40-dB signal-to-noise ratios (SNR), pitch shifting (PSH) by $\pm 4\%$, $\pm 10\%$, and $\pm 20\%$, single-echo

Table 4.2: BER (%): proposed scheme vs. other methods.

	LSB-based method [68]	**CD-based method [45]**	**FE-based method [69]**	**SSA-based method (fixed parameters) [60]**	**SSA-based method (with ad-hoc parameters) [46]**	**Proposed method**
No attack	0.00	∼0-1	0.00	0.49	0.36	0.83
Non-Malicious attacks						
G.711	0.00	∼4	0.00	0.49	0.36	1.90
G.726	51.77	∼10-25	0.00	27.66	21.07	11.12
MP3	50.49	-	-	3.69	5.39	8.67
MP4	49.53	-	-	32.79	34.19	32.52
Malicious Attacks						
BPF	50.83	-	-	50.23	50.46	21.04
AWGN (15, 40 dB)	50.70, 49.53	-	∼54	49.69, 24.53	48.67, 23.28	16.66, 9.38
PSH (−4%, −10%, −20%)	35.64, 35.33, 40.8	-	∼31, -, -	10.58, 22.03, 47.83	14.25, 36.16, 51.47	6.01, 15.57, 20.68
PSH (+4%, +10%, +20%)	34.42, 34.36, 38.03	-	-	12.44, 15.33, 20.47	7.78, 10.92, 21.94	3.51, 4.79, 8.22
Echo (20, 100 ms)	50.18, 51.34	-,∼50	-,∼5	15.76, 20.33	9.22, 18.05	4.29, 2.23
Replace (1/3, 1/2)	16.51, 24.97	-	∼57, -	17.08, 25.78	18.57, 26.25	20.07, 29.66
SCH (−4%, +4%)	49.47, 48.72	-	∼20, -	47.00, 47.19	46.58, 46.94	13.64, 13.41

addition with −6 dB, and delay times of 20 and 100 ms, substituting 1/3 and 1/2 of the watermarked signals with an un-watermarked segment, and ±4% speed changing (SCH).

The evaluation results are shown in Table 4.2. The LSB-based method showed its excellent robustness when there was no attack, but it was fragile for all other operations (except for G.711). The other methods could be considered semi-fragile and could be used for detecting tampering. However, the formant enhancement-based method was too robust when applied echo addition. It can be implied that the method cannot be used to detect tampering when a watermarked signal has been tampered with by echo addition. Our proposed method was robust in the cases of no attack and the G.711 speech coding and was fragile to other attacks. However, it was too fragile for MP4 operation. The robustness of the information hiding scheme against the G.726 speech coding was improved compared with our previously proposed method. Thus, this proposed method can be used to detect tampering in speech signals. Also, the bit-error rates of the proposed scheme can be associated with the degree of tampering,e.g., when the degree of pitch shifting increases, the BER increases. Hence, the percentage of BER can be used to indicate the degree of attacks.

4.4 Tampering Detection Ability

As described in Chapter 2 section 2.3.2, tampering can be detected by checking the mismatch between extracted-watermark bits $\hat{w}(j)$ and embedded-watermark bits $w(j)$ for $j = 1$ to M. In this section, we demonstrate how tampering detection can be done in two experiments.

First experiment, a 29×131 bitmap image of the word "APSIPA," as

shown in Fig. 4.1 (a), was used as the watermark. Consider the image size of 29×131 bitmap, which equals $3,799$ bits of information.One bit will be embedded into one frame. Thus, we need a host signal of $3,799$ frames or 95-second in length. The first 320 frames from all 12 speech signals were connected to construct a new 95-second host signal. Note that the duration was 95 seconds because our embedding capacity was 40 bps, and one frame is 25 ms in length. After the image was embedded into the host signal, we divided it into three parts, and the middle part of the watermarked signal was tampered with by performing the operations listed in Table 4.2. The reasons we choose to examine some of these operations to be tampering are as follows. Adding white noise can be considered as a distortion in the channel. Substituting watermarked speech with un-watermarked speech can be counted as a content modification. Speed changing by speeding up or slowing down a watermarked signal can be viewed as modifying the duration and tempo of speech. Pitch shifting can manipulate the individuality of the speaker. Filtering with a low pass filter is considered as removing specific frequency information of the speech.

The results of the experiment are shown in Fig. 4.1. The hidden information in the form of an image could be correctly extracted when there was no attack on the watermarked signal, as shown in Fig. 4.1 (b). The extracted hidden images from other tampered-watermarked signals are shown in Fig. 4.1 (c) to Fig. 4.1 (u). It showed that the watermark bits in the tampered part were destroyed, and the destroyed area of the extracted image was correlated with the tampered speech part. The destroyed area, in this experiment, was the middle two letters of the word "APSIPA." Moreover, the degree of tampering could be recognized from the extracted image. For example, the middle parts of the watermarked speech signals whose extracted images are shown in Fig. 4.1 (n) and Fig. 4.1 (s) were destroyed by adding white Gaussian noise (AWGN). It can be recognized that the middle part of the extracted image of Fig. 4.1 (s) was more severely damaged because the stronger noise was added to the speech signal of Fig. 4.1 (s). Similarly, Fig. 4.1 (g), Fig. 4.1 (l), Fig. 4.1 (q), Fig. 4.1 (h), Fig. 4.1 (m), and Fig. 4.1 (r) where all of them were attacked by pitch shifting with different degrees showed the same trend. The middle part of the extracted image was more severely destroyed when the degree of the attack was raised. Therefore, we can use the destroyed areas and their characteristics to identify the tampered parts of the watermarked signals and the degree of tampering.

In addition to the tampered position and the tampering degree, we could approximately predict the tampering type by analyzing the damaged area of the extracted image. A singular spectrum is maintained when the embedding watermark bit was 0 regarding our embedding rule. Therefore, if the damaged

area is dark, such as those in Fig. 4.1 (p) and Fig. 4.1 (u), that area is possible be extracted from an substituted un-watermarked segment. The reason is because a singular spectrum is typically convex, and singular values between $\sqrt{\lambda_p}$ and $\sqrt{\lambda_q}$ are therefore below the straight line that connects $\sqrt{\lambda_p}$ and $\sqrt{\lambda_q}$. Hence the extracted bit is 0, represented by the black pixel.

As mentioned in subsection 3.4.1, removing high-frequency components from a signal can refer to reduce its high-order singular values. Therefore, removing high-frequency components increasing the chance to obtain a watermark bit 0 when decoding the watermark bit. Consequently, the destroyed area of the extracted image got darker, as evidenced in Fig. 4.1 (l) and Fig. 4.1 (g), when the pitches of the middle speech parts were decreased by 10% and 20%, respectively. In contrast, adding high-frequency components can provoke high-order singular values to increase in value.

The second experiment, the attack was simulated by using a vocoder named STRAIGHT [6]. For example, we can use STRAIGHT to revise the sentence "No, I did not" to become"Yes, I did" by replacing "No" with "Yes" and then removing "not" from the sentence. The simulation steps are as follows. First, a watermark, which is a 166×23 bitmap image of the word "STRAIGHT," was embedded into a host signal of 96-seconds long. Figure 4.2 (a) showed an extracted image with no attack on the watermarked signal. Second, STRAIGHT read the watermarked signal to get specific features: the fundamental frequency (F0), aperiodic information, and an F0 adaptively smoothed spectrogram. These extracted specific features were used to synthesize another speech signal in third step, and final step, the synthesized speech signal was then substituted the watermarked signal in the second half. It can be seen that a replaced part can change critical information in the host signal and mislead the listeners. Fourth, the modified signal obtained from the previous step was inputted into the extraction process to get the watermark. The extracted watermark is shown in Fig. 4.2 (b). Note that the extracted watermark of the replaced segment was damaged. Note that this experiment gives similar results as the first experiment, our scheme could be used to recognize a tampered segment in a speech signal. The substituting part of a speech signal with a synthesized signal is different from substituting it with an un-watermarked part since the synthesized signal has different distortion. For example, the SDR of the synthesized speech signal was -27.81 dB, which is considerably low. Therefore, a synthesized signal can be viewed as a noisy speech signal. Hence, the damaged area in Fig. 4.2 (b) looks similar to that shown in Fig. 4.1 (s).

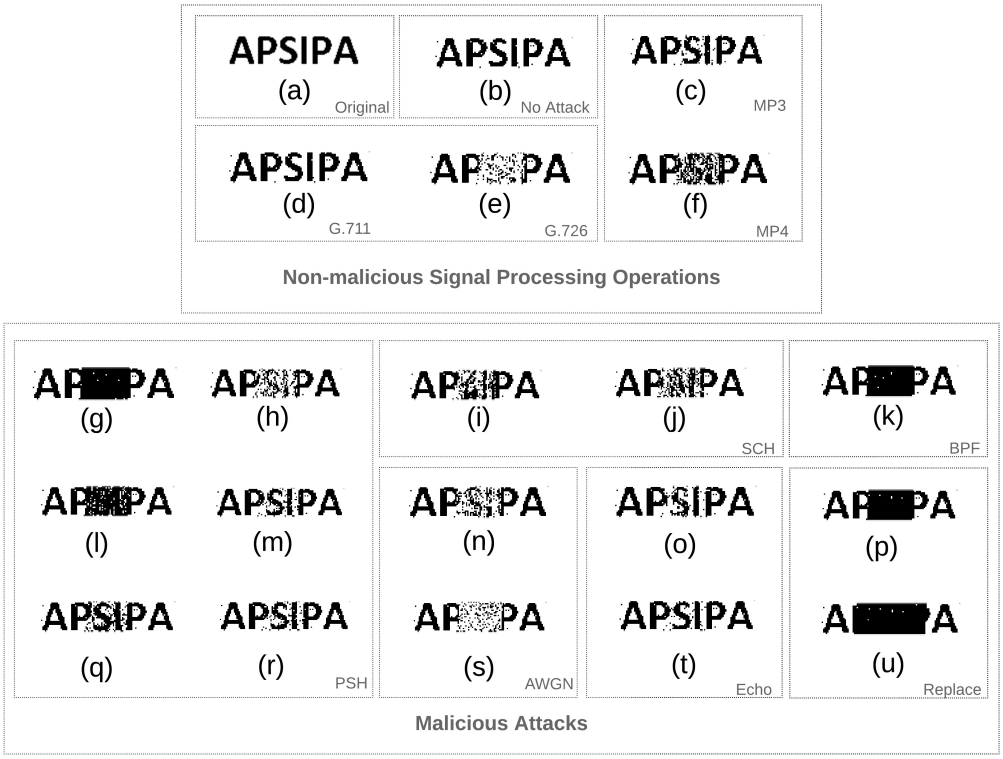


Figure 4.1: Comparison of watermark image between original image (a) and reconstructed images after performing following signal-processing operations: (b) no attacks, (c) MP3, (d) G.711, (e) G.726, (f) MP4, (g) PSH −20%, (h) PSH +20%, (i) SCH +4%, (j) SCH −4%, (k) BPF, (l) PSH −10%, (m) PSH +10%, (n) AWGN (40 dB), (o) echo (100 ms), (p) replace (1/3), (q) PSH −4%, (r) PSH +4%, (s) AWGN (15 dB), (t) echo (20 ms, and (u) replace (1/2).

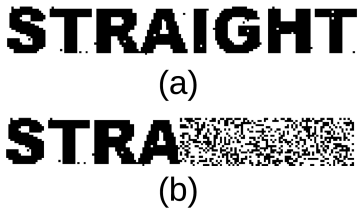


Figure 4.2: Comparison of extracted watermark-image: (a) no attacks and (b) second half of speech signal substituted by synthesized speech signal.

Table 4.3: An accuracy of tampering detection of the proposed method

	FRR	FAR
Accuracy	5%	25%

4.4.1 Tampering detection accuracy

The SSA-based AIH is designed for tampering detection. The detection performance was measured by false rejected rate (FRR), i.e., false negative (FN) and false acceptance rate (FAR), i.e., false positive (FP). Firstly, speech signals will be judged as tampered or non-tampered using BER. If the BER of the signal is higher than 10 %, the signal will be judged as tampered and if lower than 10 % is non-tampered. The incorrect judgment for each signal belonged to one of the following two categories FRR and FAR. FRR is defined as the intact signals that were judged as tampered and FRR were calculated as follow.

$$\text{FRR} = \frac{N_{FRR}}{N} \times 100\%, \tag{4.3}$$

where N_{FRR} is numbers of signal in FRR categories, and N is the total number of tested signals.

FAR is a tampered signal judged as non-tampered and FAR were calculated as follow.

$$\text{FAR} = \frac{N_{FAR}}{N} \times 100\%, \tag{4.4}$$

where N_{FAR} is numbers of signal in FAR categories, and N is the total number of tested signals.

An accuracy of tampering detection of the proposed method is shown in Table 4.3.

It can be seen that FRR is 5% because our proposed method is not robust to MP4 operation. The robustness of MP4 of our proposed method had BER at 32.52%. All speech signals performed MP4 were judged by the tampering detection system as tampered speech. For FAR rate is 25%, this false rejected rate come from the low degree attacks. Let consider BER for each attack in Table 4.2. The low degree attacks have BER lower than 10% were judged as non-tampered. Actually, this FAR rate does not reflect the facts because some low degree does not affect the sound quality or robustness. We should define a threshold of the attacks carefully. Table 4.4 showed the accuracy of the system after the threshold of attacks is defined. Here we

Table 4.4: An accuracy of tampering detection after defined threshold of attacks

	FRR	FAR
FE-based method [56]	1.71 %	3.65 %
Proposed method	5.00 %	12.50 %

remove the low degree of attacks by referencing the recent paper that detects tampering in speech signals [56]. The Adding white Gaussian-noise with a high signal-to-noise ratio and the tampered with a low degree of pitching shifting were removed. It can be seen that now the FAR rate only causes by echo adding. The false acceptance was decreased from 25% to be 12.5%. However, the lower percentage of FRR and FAR, the better in performance of the detection system. Therefore, robustness to MP4 of the proposed scheme should be improved as well as the scheme should be more fragile to echo adding in order to improve tampering detection.

4.5 Computational Time

The DE-based parameter estimation's computational time is significantly high because of many simulated processes, i.e., an embedding process, an extraction process, and many attacks simulation. As a consequence, SVD was performed many times for each input segment, and SVD is time-consuming. Also, the search space of DE is enormous to cover all possible value. The computational time is significantly reduced using CNN-based parameter estimation instead of DE-based estimation in the information hiding scheme. A 10-fold cross-validation was conducted to assure model stability. All of the simulations have been experimented with a personal computer with Windows 10 (Home Edition). The CPU was a 7^{th} generation Intel® Core™ i5-7360U with 2.3 GHz clock speed of and 8 GB memory size of a 2,133 MHz speed. A comparison of computational times of both methods is shown in Table 4.5. Note that the CNN-based method was approximately 2 million times faster than that of the DE-based method.

Even the CNN-based parameter estimation is very helpful to reduce the computational time, but we have to trade-off with the accuracy of the parameter estimation. We compare parameters obtained from the DE-based method and the parameters obtained from the CNN-based method with the root-mean-square error (RMSE). Figure 4.3 showed a comparison of parame-

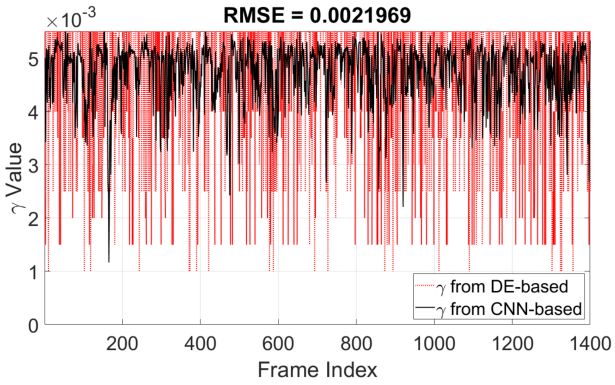
Table 4.5: Comparison of computational times for of parameter estimation of the method based on differential evolution and the method based on CNN.

	Computational Time	
	time/frame	**time/signal**
DE-based method	6 minutes	32 hours
CNN-based method	0.195 millisecond	0.065 second

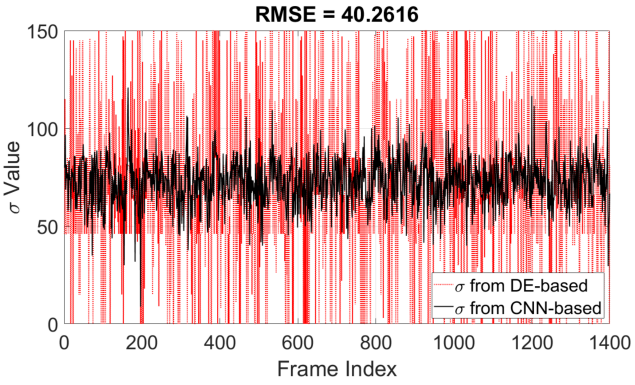
Table 4.6: Comparison of robustness and inaudibility of scheme when automatic parameterization is based on differential evolution and when it is based on CNN.

	DE-based Method	**CNN-based Method**
BER_{NA} (%)	0.00	0.83
BER_{G711} (%)	0.00	1.90
BER_{G726} (%)	25.00	11.12
BER_{MP3} (%)	10.00	8.67
BER_{MP4} (%)	30.00	32.52
LSD (dB)	0.71	0.45
SDR (dB)	30.63	35.51

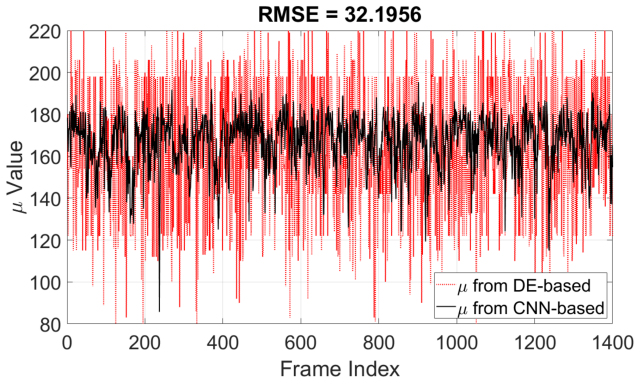
ters obtained from the DE-based method and the parameters obtained from the CNN-based method. The RMSE of each parameter estimation are as follows, RMSE of parameter γ was 0.0022, the RMSE of parameter μ was 32.1956, and the average RMSE of parameter σ was 40.2616. The RMSE values of the parameters μ and σ seems to be quite large. However, if we consider on the robustness and inaudibility of the scheme when both methods were applied were comparable, as shown in Table 4.6. Figure 4.4. showed an example of a singular spectrum of a frame that is embedded with parameters estimated from the DE-based method and those estimated with the CNN-based method. In this experiment, the error (or difference) between the two parameter vectors $[\mu_{\text{DE}} \ \ \sigma_{\text{DE}}]^{\text{T}}$ and $[\mu_{\text{CNN}} \ \ \sigma_{\text{CNN}}]^{\text{T}}$ was $\sqrt{(\mu_{\text{DE}}-\mu_{\text{CNN}})^2+(\sigma_{\text{DE}}-\sigma_{\text{CNN}})^2} = 90.56$, may seem to be large compared with the RMSE, but the modified singular spectra of both estimation do not look much different.



(a)



(b)



(c)

Figure 4.3: RMSE of γ, μ, and σ from DE-based parameter estimation and CNN-based parameter estimation.

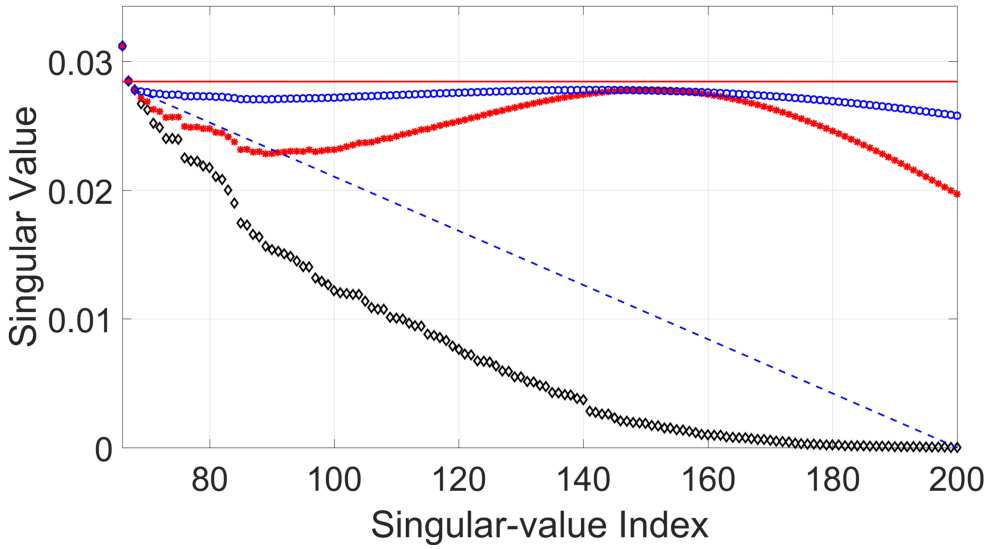


Figure 4.4: Example of singular spectrum of embedded frame. "$\diamond$" denotes original singular spectrum, "$*$" denotes modified singular spectrum where parameters are obtained from CNN-based method, and "$\circ$" denotes singular spectrum where parameters are obtained from DE-based method, red solid line denotes γ threshold, and dashed line denotes a straight line connected the first and last singular value to be modified.

4.6 Discussion

The information hiding scheme for tampering detection is constructed, and the scheme can be reducing the computational time of parameter estimation. Two main points should be discussed: the CNN-based method effectiveness and the detection accuracy of the system. The first point is related to the CNN-based method effectiveness. The CNN-based method effectiveness is discussed in three aspects: the effectiveness of the CNN-based method, the role of the cost function for DE optimizer, and the computational time calculation. The second point is related to the detection accuracy of the system and its effectiveness.

Let us start with the effectiveness of the CNN-based method. The computational time of parameter estimation is significantly reduced. However, the effectiveness of the CNN-based method cannot go beyond that of the DE-based method since DE is used as the basis of the framework that we use to generate the training dataset. The performance of the CNN-based method is typically poorer than the DE-based method because there is an error in the learning (or fitting) process during the building of the CNN in most cases. A crucial factor that is responsible for the effectiveness of the DE algorithm is the cost function. In this work, the cost function and some DE hyper-parameters, such as the upper bounds and the lower bounds of the parameters, play an essential role in balancing robustness and inaudibility. In other words, the CNN-based method's effectiveness depends on how well supervised by the training dataset provides by the DE-based method. The better cost functions provide a better quality of the training dataset.

Here the role of the cost function is discussed. Defining a good cost function is not trivial, and it is presumably impossible to explore all possible cost functions. The basic assumption applied to setting cost function is that the cost function should include two terms: one representing robustness and the other representing inaudibility. The eight different settings was used, as shown in Table 4.7. Evaluations of the robustness and inaudibility when these cost functions were used in the DE optimizer are shown in Table 4.8. Note that these functions were evaluated by using only 40 frames due to the expensive computational cost of DE.

Cost functions C_1 and C_2 look similar. Both take the LSD into account and equally weigh the terms representing inaudibility and robustness equally. Also, they assign the same weight β_i for the same BER conditions. The only difference is the upper bound of γ, i.e., the search space of γ of C_2 is smaller than that of C_1. We found that their average BERs were comparable, but C_1 yielded a better sound quality. Therefore, we can safely infer that we can use the possible range of γ to control the sound quality of a watermarked

Table 4.7: Eight cost functions studied in our investigation.

Cost function	Hyperparameter	
	Lower bound of γ	Upper bound of γ
$C_1 = \sqrt{\mathrm{LSD}^2 + \overline{\mathrm{BER}}^2}$, where $\overline{\mathrm{BER}} = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.001	0.015
$C_2 = \sqrt{\mathrm{LSD}^2 + \overline{\mathrm{BER}}^2}$, where $\overline{\mathrm{BER}} = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.007	0.015
$C_3 = \sqrt{\frac{2}{10}\mathrm{LSD}^2 + \frac{8}{10}\overline{\mathrm{BER}}^2}$, where $\overline{\mathrm{BER}} = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.007	0.015
$C_4 = \sqrt{\frac{3}{10}\mathrm{LSD}^2 + \frac{7}{10}\overline{\mathrm{BER}}^2}$, where $\overline{\mathrm{BER}} = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.007	0.015
$C_5 = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.001	0.015
$C_6 = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.007	0.015
$C_7 = \frac{1}{3}\mathrm{BER}_{\mathrm{NA}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP3}} + \frac{4}{21}\mathrm{BER}_{\mathrm{MP4}} + \frac{4}{21}\mathrm{BER}_{\mathrm{G711}} + \frac{2}{21}\mathrm{BER}_{\mathrm{G726}}$	0.001	0.0085
$C_8 = \mathrm{BER}_{\mathrm{NA}}$	0.007	0.015

signal.

Let us consider C_2 and C_3. For this pair of cost functions, we wanted to investigate the outcome when we adjusted the weights between the robustness term ($\overline{\mathrm{BER}}$) and the inaudibility term (LSD). In C_3, the robustness was weighted three times greater than the inaudibility. We expected that DE with C_3 would favor robustness much more than inaudibility. However, the results showed that the average BER of C_3 was about 25% less than that of C_2, whereas the LSD of C_3 was about 50% greater than that of C_2.

Similarly, when we considered the outcomes of C_2, C_3, and C_4 together, we found that controlling the balance between robustness and inaudibility by adjusting the weight between the LSD and the $\overline{\mathrm{BER}}$ was not effective, as evidenced in Table 4.8. Thus, we tried another strategy, i.e., we used the size of the search space of γ to control the sound quality.

Let us consider the outcomes of C_5, C_6, and C_7 in comparison with C_2, C_3, and C_4. It can be seen that, when we set the upper bound of γ appropriately, we could gain an improvement in sound quality while the BER level was maintained.

Finding an efficient cost function is not the primary focus of this work, but it is of importance due to the fact that it will help us to generate a better training dataset for the CNNs. Therefore, in this work, the robustness against critical speech signal operations such as G.711, G.726, MP3, and MP4 should be ensured. DE optimizer can simulate this attack to find the parameter that makes the scheme robust to these operations. Also, adding

Table 4.8: Evaluations of robustness and inaudibility when different cost functions were deployed.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
BER_{NA} (%)	10.00	17.50	7.50	12.50	0.00	0.00	0.00	0.00
BER_{G711} (%)	10.00	17.50	7.50	12.50	0.00	0.00	0.00	0.00
BER_{G726} (%)	47.50	42.50	37.50	45.00	10.00	20.00	15.00	25.00
BER_{MP3} (%)	22.50	27.50	17.50	27.50	2.50	2.50	2.50	10.00
BER_{MP4} (%)	40.00	32.50	30.00	32.50	7.50	7.50	5.00	30.00
LSD (dB)	0.12	0.19	0.30	0.23	0.70	0.79	0.50	0.71
SDR (dB)	58.77	47.99	39.36	47.75	27.73	27.08	35.17	30.63

more signal processing operations into the DE optimizer could provide the training dataset with high robustness. We will tackle this problem in the future.

The last issue for CNN-based method effectiveness is the computational time calculation. A comparison of computational times of both methods is shown in Table 4.5. Note that the computational time calculation of the CNN-based method excludes the training phase. The time in the training phase of CNN is excluded from the computational time of the CNN-based method because the model was trained only one time, and the model was used to estimate the parameter multiple times, then the training time can be discarded. Moreover, if the CNN is well trained, then the CNN can estimate the embedding parameter of unknown speech without retraining. This idea was examined by testing the CNN-based parameter with 37.5 % of the untrained speech signal. The result of extraction precision from the 37.5 untrained testing set was comparable with the extraction precision of all testing set were trained. This result motivates that the CNN-based parameter can estimate the embedding parameter of unknown speech without retraining. However, since the experiment was set up for concept checking and was not checking with a large dataset, we only imply that the training time can be excluded from computational time calculation. The estimation of embedding parameters of unknown speech will be considered in future work.

The second point we want to discuss is related to the accuracy of tampering detection. As we have seen, the result of tampering detection accuracy is shown in Table 4.4. We remove the low degree of attacks, such as adding white Gaussian noise with a high signal-to-noise ratio and tampered with a low degree of pitching shifting, but the FRR and FAR are still high as 5 % and 12.5 %, respectively. It isn't easy to compare our proposed method

with another method since there are no standards on which type of signal operation to be tested as non-malicious attacks nor which type and degree of attacks to be tested as malicious attacks. Some tapering detection focused only on content tamperings, such as insertion or deletion [52]. The signal operation and malicious attacks were tested on cochlear delay-based method [45] and formant enhancement-based method [56], and SSA-based method [73] are also different. Almost all method regarding tampering detection showed the result in BER for different type of attacks, but there is only formant enhancement-based method [56] that show the result on tampering detection accuracy. Thus, we compare our proposed method with FE-based method [56]. Note that the embedding capacity of the proposed method is 40 bps while the embedding capacity of the FE-based method is 8 bps.

The lower percentage in FRR and FAR refers to the better performance of the detection system. The proposed method has FRR at 5 % because the scheme is not robust to the MP4 operation. The FE-based method has FRR at 1.71 %, but the MP4 operation does not include the FE-based method experiment. The reason that the proposed method does not become robust to MP4 because we hide the secret information in high order singular value that implies a high-frequency oscillator component of the speech signal, which hidden information could be lost on MP4 compression. The proposed method has FAR at 12.5 % because the scheme is not fragile to echo adding. The FE-based method has FAR at 3.65 %, but the echo adding does not include the FE-based method experiment. The invariance property of singular value may cause the proposed method not to become fragile to echo adding. However, we cannot say that our proposed method successful in tampering detection. The number 5 % of FRR may be annoying the user that the system does not accept the MP4 operation as the non-tampered speech. When we consider the false negative (FN), or FRR, there is less critical than false positive (FP) or FAR because FAR refers to the system accept the tampered speech as the non-tamper one. The FAR of our system is 12.5 %. If this system installed to discard the tampered speech in a sensitive system such as a banking system, this could lead to a problem. Therefore, robustness to MP4 of the proposed scheme should be improved as well as the scheme should be more fragile to echo adding in order to improve the effectiveness of tampering detection.

4.7 Summary

In this chapter, we propose the core structure of SSA-based AIH for tampering detection. The main require property of the information hiding scheme for tampering detection is semi-fragile, i.e., the information hiding

scheme should be robust to non-malicious attacks and fragile to the attacks. This proposed scheme is base on a singular SSA-based information hiding method. Hence, a watermark was embedded into a host speech signal with the same concept in previous: modifying a part of its singular spectra. As we discover that the modification affects the sound quality and robustness of the scheme, it can also make the scheme robust, fragile, or semi-fragile. Therefore, in this work, the part of the singular spectrum to be modified must be carefully selected to make the scheme semi-fragile for tampering detection. Previously, we found that a DE algorithm can be deployed to select the appropriate part for modification, but it was costly in computational time. In this work, CNN-based parameter estimation is offered to replace DE. However, DE was used as the basis of a framework for generating a high-quality dataset for CNN training. The results from the experiment showed that the scheme deployed CNN-based parameter estimation could correctly detect whether tampering occurs or not, and it could locate tampered areas and roughly predict the types and degrees of tampering. When using CNN-based parameter estimation, the computational time could reduce by approximately 2 million times and improve the watermarked signal's sound quality. In addition, the information hiding scheme is entirely blind because the estimation can be used to find the parameters in both the embedding and extraction processes. However, the proposed method needs to be improved in the tampering detection accuracy because the scheme has FRR at 5 % and FAR at 12.5 %. This error comes from the scheme is not robust to MP4 operation, and the scheme is not fragile to echo adding. Therefore, robustness to MP4 of the proposed scheme should be improved as well as the scheme should be more fragile to echo adding in order to improve the effectiveness of tampering detection.

Chapter 5

Application of Information Hiding

There are two application shown in this chapter.

The first application is related to the last example in the introduction chapter. The situation concerns two people who want to communicate secretly, but they do not want others to know that they are privately communicating. In some companies, they monitor the email, song, recording that send through the network. Their policy does not allow the employees to encrypt the message, including a song or sound recording, since they cannot monitor those encrypted messages. In this application, the singular spectrum analysis (SSA)-based information hiding method with the transformation method to provide a secret and secure channel on speech signals. The SSA-based AIH will provide secret channel and Arnold transformation make the secret channel to be secured. Here we want to define clearly the difference between a secret and secure channel. A secret channel focuses on how difficult to know the existence of the channel, while a secure channel focuses on how difficult to access data on the channel.

The second application is related to the situation that the company distributes the message to everyone, but the employees at the different levels can only access the message related to their authorization. In this application our information hiding method can be deployed encryption method to provide the accessing data at the different levels.

In summary, there are two different applications in this chapter. The first application provides the secret and secure channel, while the second application provides the accessing data at the different levels. Both methods deploy SSA-based AIH and Arnold transformation. The concept and detail of applying Arnold transformation and SSA-based AIH to solve different problems will be described in this chapter. The evaluation and result also be provided.

5.1 Statement of the problem

There are two different problems in this chapter. The first problem is related to the last example from the introduction chapter. The situation concerns two people who want to communicate secretly, but they do not want others to know that they are privately communicating. In some companies, they monitor the email, song, recording that send through the network. Their policy does not allow the employees to encrypt the message, including a song or sound recording, since they cannot monitor those encrypted messages. The second problem is related to the situation that the company distributes the message to everyone, but the employees at the different levels can only access the message related to their authorization. Note that encryption is not prohibited since the company distributes the message.

Both methods use an SSA-based AIH core structure to solve this problem and adopt Arnold's transformation differently to solve each question. From literature of encryption algorithm, RSA, a widely used public-key cryptography system, has been used for speech data encryption and decryption, but it is limited by the maximum number of signals that can be encrypted at a single time [74]. The chaotic algorithm had a shorter computation time, but there was a trade-off with security level [75]. Multiple scrambling was applied to strengthen information hiding, but the speech contents could be accessed by anyone [76]. The audio encryption algorithm using an elliptical curve and Arnold transformation was evaluated to determine its suitability for information hiding, but it did not include the implementation or evaluation of an information hiding scheme [77]. The hybrid domain was applied in audio watermarking with chaotic encryption, but the encryption was only applied to the watermark signal [78]. The Arnold transformation performs on a matrix and our SSA-based method and analyze the singular value of the matrix represent the signal. Thus, Arnold transformation is suitable to use for SSA-based AIH.

5.2 Background

5.2.1 Singular spectrum analysis-based Information hiding

SSA-based information hiding is the same as the one was introduced in Chapter 3. The scheme used the concept of basic SSA to analyze host signals and extract the singular spectra, and the watermark signal was hidden in a part of the spectra. However, there is a bit of difference between this

proposed and the one introduced in Chapter 3, i.e., the rule for embedding the watermark bit and the method to make a secured watermark signal.

Therefore, to recall the SSA-based AIH concept, we provide a short brief of SSA-based AIH, which has two main processes: embedding and extraction.

In embedding, the host speech signal is segmented into non-overlapping frames. One watermark bit is embedded into one frame. Thus, the number of frames is equal to the number of the watermark bits to be embedded. Then the trajectory matrix $\mathbf{F}$ which represents each frame F is constructed. Singular value decomposition (SVD) is performed on each trajectory matrix $\mathbf{F}$ to obtain each frame's singular spectra. The singular spectra are modified to hide the watermark bit (0 or 1), and the part of the singular spectra to be modified depends on the requirement of the information hiding application. The modified trajectory matrix $\mathbf{Y}$ is constructed by SVD reversion and then hankelized. The hankelization of a modified trajectory matrix $\mathbf{Y}$ yields a signal G, where G is a frame of the watermarked signal. The frames are stacked to reconstruct the watermarked signal.

In extraction, the watermarked signal is first segmented into non-overlapping frames, and the trajectory matrix is constructed in the same way as in the embedding process. Then SVD is performed on the trajectory matrix to obtain the singular spectra. The singular spectra of the signal of each frame are typically convex; however, the watermark bit embedded into an interval of the singular spectrum of a host frame results in a concave part on the interval of the singular spectrum of the reconstructed, watermarked frame. This property can be utilized to extract the watermark bit from each frame.

5.2.2 Arnold scrambling algorithm

The Arnold scrambling algorithm, or Arnold transformation, describes a discrete mapping from site (x_t,y_t) to site (x_{t+1},y_{t+1}) with circumference N, where ($0 \leq t < N$) and mod is a modulo function.

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \bmod N. \tag{5.1}$$

Arnold transformation is used to alter a matrix X of dimension NxN into a matrix $X^{'}$ to decrease the correlation coefficient between the matrices. Arnold transformation is cyclical, and iterated. The scrambling key is needed as a secret key to identify the number of iterations during the transformation process to bring back the original matrix. In the proposed method, Arnold transformation is applied to the watermark signal to provide a secured watermark signal, and it is, in turn, applied to the watermarked signal for encryption.

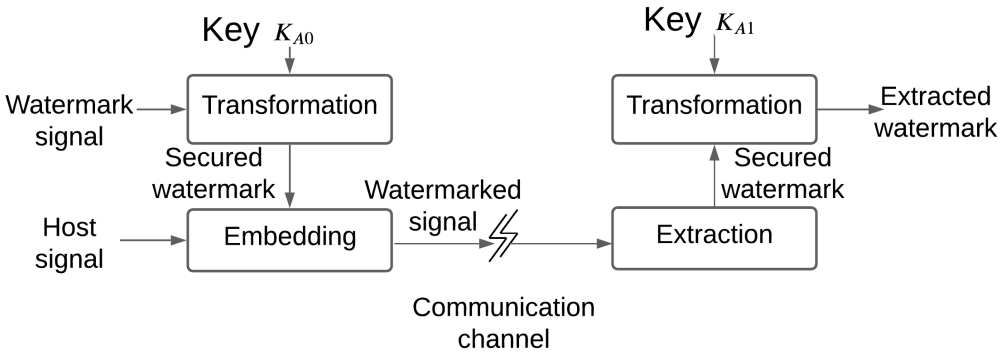


Figure 5.1: Secret and secure channel: Emitter (left), and receiver (right).

5.3 Proposed method

This section introduces two systems of information hiding adopted transformation.

5.3.1 The scheme for construct secret and secure channel

The first scheme is constructed for solving the first situation that two people want to communicate secretly. The scheme consists of the emitter side and the receiver side, as illustrated in Figure 5.1. The watermark signal is transformed using key K_{A0} to obtain the secure watermark and embedded into the host signal to produce a watermarked signal. The watermarked signal is then sent through the communication channel on the emitter side. The watermarked signal is received and decoded to obtain the secured watermark, which is later transformed using K_{A1} to obtain the original watermark on the receiver side. The watermark or hidden information is made to be secured in this scheme and sent through the network directly. The hidden information does not attract listeners' attention then only an authorized person with the correct key can access hidden information.

5.3.2 The scheme deployed encryption

The second scheme is constructed to solve the second situation that the company distributes the message to all, and access is granted depending upon employees' level. The scheme consists of the emitter side and the receiver

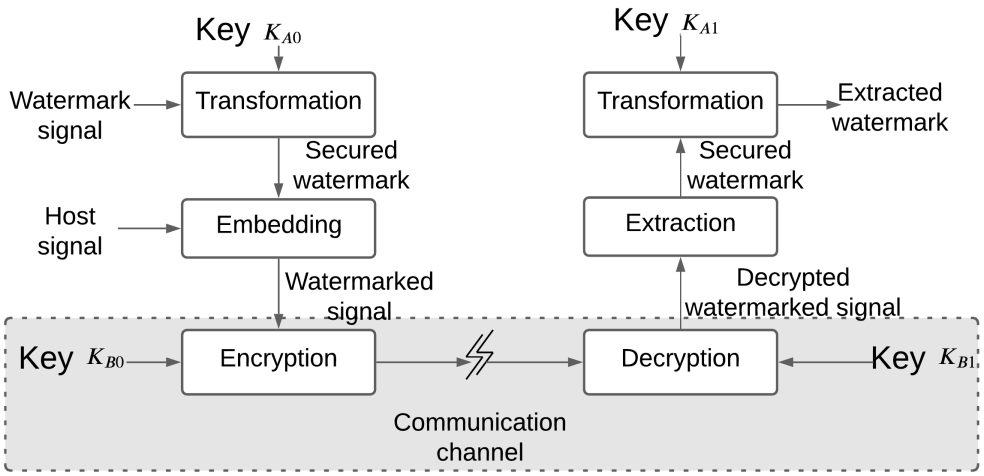


Figure 5.2: Scheme deployed encryption: Emitter (left), and receiver (right).

side, as illustrated in Fig. 5.2. The watermark signal is transformed using key K_{A0} and embedded into the host signal to produce a watermarked signal, which is later encrypted using K_{B0} to be sent through the communication channel on the emitter side. The watermarked signal is decrypted using key K_{B1} and then decoded to obtain the secured watermark, which is later transformed using K_{A1} to obtain the original watermark on the receiver side. Since these two schemes share the same basic structure and only the shade area in Figure. 5.2 is different, therefore, we will explain in detail of emitter side and receiver of both schemes and point out the different points.

The details of each side are as follows

5.3.3 Emitter side

A detailed diagram of the emitter side in the proposed scheme is shown in Figure 5.3. The left-hand side of Figure 5.3 (a) shows the procedure for creating a secured watermark. The right-hand side (b) shows the procedure for inserting secured watermarks into the host signal to obtain the watermarked signal and encrypting the watermarked signal. The watermark signal W is divided into vectors converted to NxN matrix $\mathbf{W}$, and Arnold transformation alters the watermark matrix $\mathbf{W}$ to obtain altered watermark matrix $\mathbf{W}^{'}$ using key K_{A0}, where key K_{A0} is a pre-defined number of transformation iterations. Next, the altered watermark matrix $\mathbf{W}^{'}$ is converted to a secured watermark signal $W^{'}$ to be embedded into the host speech signal. The confidentiality of

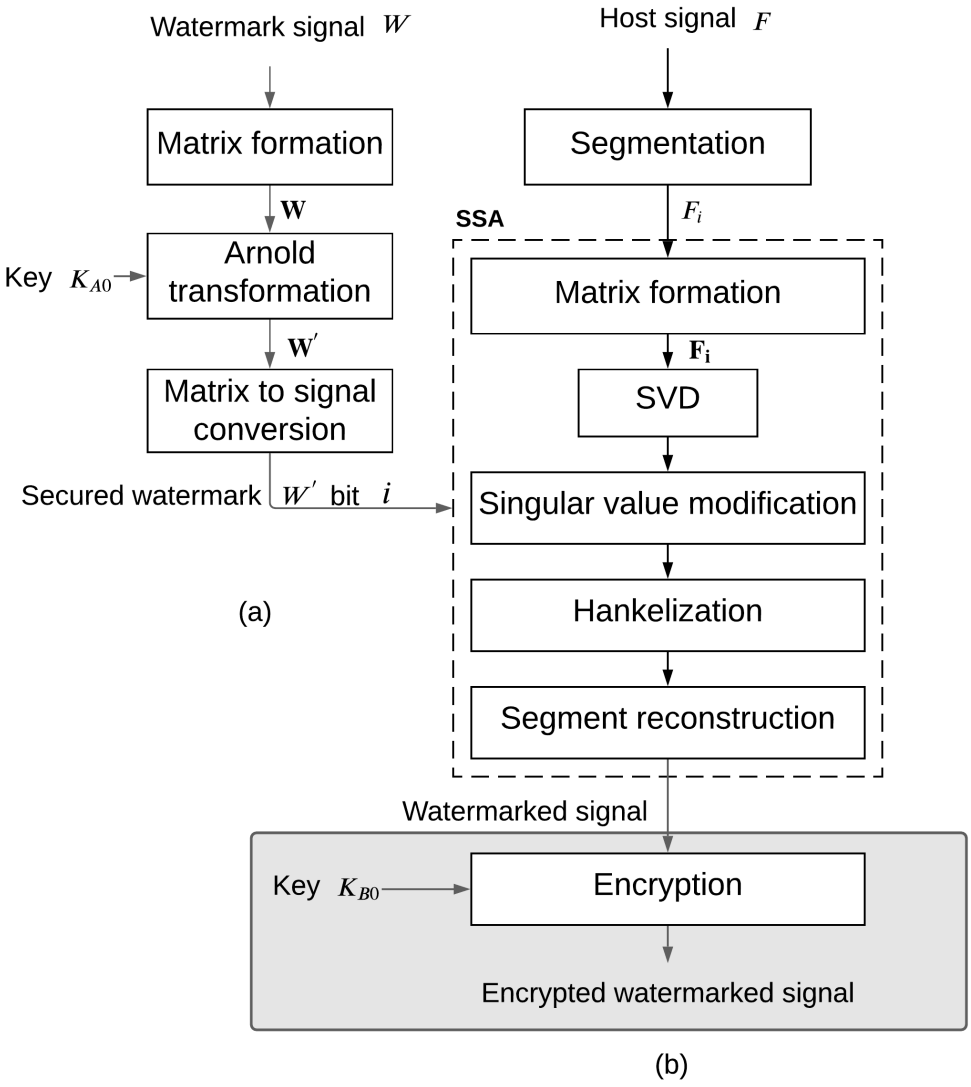


Figure 5.3: Emitter side.

the watermark signal can be strengthened as a result.

The method of building a secret and secure channel sends the watermark directly through the network, and it does not apply the encryption step, as shown in the shaded area of the left-hand side of Figure 5.3. In contrast, the method to provide different accessing data requires the encryption step, and finally, the encrypted watermarked signal is sent through the network.

The method of embedding process to obtain a watermarked signal (for the secret and secure scheme) or obtain an encrypted watermarked signal (for different granting access) are as follows.

1. *Segmentation.* The host speech signal is segmented into frames of equal length M, where M is the total number of samples in each frame.

2. *Matrix formation.* A signal F of each frame is mapped to a trajectory matrix $\mathbf{F}$ of the size $L \times K$, where $F = [f_0\ f_1 \ldots f_{M-1}]^{\mathrm{T}}$ where f_i for $i = 0$ to $M-1$. The signal F is mapped to matrix $\mathbf{F}$ by the following relation

$$\mathbf{F} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{K-1} \\ f_1 & f_2 & \cdots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \cdots & f_{M-1} \end{bmatrix}. \tag{5.2}$$

   where L is a *window length*, and $2 \leq L \leq M$, and K is $M-L+1$.

3. *Singular Value Modification.* A singular spectrum is modified on the basis of the secured watermark bit to be embedded. Given a singular spectrum $\{\sqrt{\lambda_0}, \sqrt{\lambda_1}, ..., \sqrt{\lambda_q}\}$, a specific part of this singular spectrum, which is $\{\sqrt{\lambda_p}, \sqrt{\lambda_{p+1}}, ..., \sqrt{\lambda_q}\}$, is modified on the basis of the secured watermark bit w with

$$\sqrt{\lambda_i^*} = \begin{cases} \sqrt{\lambda_i} + \alpha_i(\sqrt{\lambda_p} - \sqrt{\lambda_i}), & \text{if } w = 1, \\ \sqrt{\lambda_i} \quad \text{(i.e., unchanged)}, & \text{if } w = 0, \end{cases} \tag{5.3}$$

   where $\sqrt{\lambda_i^*}$ is the modified singular value for $i = p$ to q, $\sqrt{\lambda_p}$ is the largest singular value that is less than $\gamma \cdot \sqrt{\lambda_0}$, α_i is an embedding strength, as defined in [46]. Note that γ is a pre-defined value to control the number of singular values to be modified.

4. *Hankelization.* A watermarked matrix $\mathbf{X}^*$ is computed as the product of $\mathbf{U}\mathbf{\Sigma}^*\mathbf{V}^{\mathrm{T}}$ and then hankelized to obtain the signal F^*, which is the watermarked segment. The hankelization is the average of the anti-diagonal $i+j=k+1$, where i and j are the row index and the column index, respectively, of an element of $\mathbf{X}^*$, and k (for $k=0$ to $M-1$) is the index of element F^*.

5. *Segment Reconstruction.* The watermarked signal is finally produced by sequentially concatenating all watermarked segments.

6. *Encryption.* The watermarked signal from the previous step is transformed into an NxN matrix and encrypted using key K_{B0} to scramble its elements. The encrypted matrix is reshaped into one dimension resulting in an encrypted watermarked signal to be sent through the communication channel.

Note that the Arnold transformation was applied to secure a watermark signal and to encrypt the watermarked signal. However, the process was referred to as a *transformation* when performed on a watermark signal, and *encryption* when performed on the encrypted watermarked signal. This is to clarify which signal is being transformed as each process differs slightly. For example, the matrix size NxN of the watermark signal may differ from that of the encrypted watermarked signal due to the signals' size difference. NxN only represents the square matrix, and its value N can be pre-defined. Since the matrix sizes differ, matrix construction and signal reconstruction differ as well.

5.3.4 Receiver side

A detailed diagram of the receiver side in the proposed scheme is illustrated in Figure 5.4. There are two main procedures on the receiver side, (a) extracting a secured watermark, and (b) retrieving the original watermark.

There is a difference in the receiver of these two schemes where the method of building a secret and secure channel receive the watermarked signal directly from the network, so it does not require the decryption step, as shown in the shaded area of the left-hand side of Figure 5.4. In contrast, the method to provide different accessing data requires the decryption step because it receives the encrypted watermarked signal from the network.

The left-hand side of Figure 5.4 shows the five steps in extracting a secured watermark. The first step is *Decryption.* The received signal is reshaped into NxN matrix and is decrypted using key K_{B1} to produce the watermarked signal. Note that key K_{B1} on the receiver side matches K_{B0} on the emitter

side. The decrypted watermarked signal is then passed through the next three steps, which are *Segmentation*, *Matrix formation*, and *SVD*, as is done on the emitter side. The last step is *Decoding the singular spectra*. The secured watermark bits are extracted by decoding the singular spectra, and how the spectra are decoded depends on how they are modified in the embedding process. The embedding rule in equation (5.3) results in the concave part on the singular spectra if embedding bit 1. Thus, by this property, the secured watermark bit is extracted.

The right-hand side of Figure 5.4 (b) shows how the secured watermark is transformed to obtain the original watermark again. The secured watermark signal is divided and converted to an NxN matrix. The Arnold transformation transforms the secured watermark matrix using key K_{A1} to recover the original watermark. Note that key K_{A1} on the receiver side matches of K_{A0} on the emitter side.

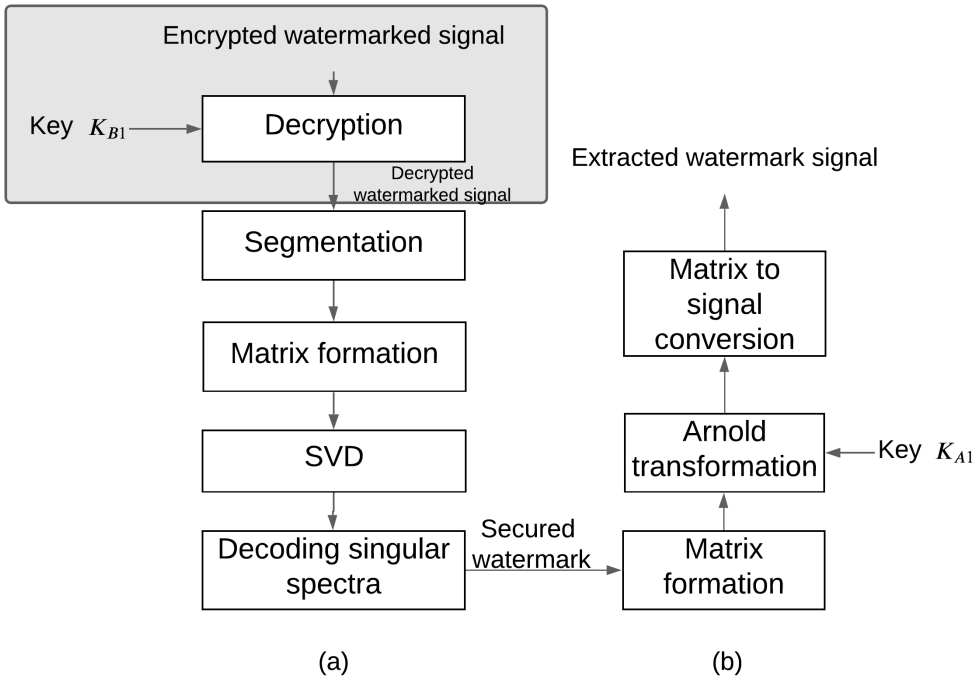


Figure 5.4: Receiver side.

5.4 Evaluations and results

In the experiment, twelve speech signals of Japanese sentences uttered by six men and six women from the ATR database (B set) were used [70]. The speech signals were one-channel with a 16-kHz sampling rate and 16-bit quantization. To evaluate the scheme used for building secret and secure channels, we focus on how the scheme successfully achieves imperceptible property because this property is related to the attention of listeners and attackers. Moreover, the accuracy of watermark extraction without the correct key is considered. To evaluate the scheme deployed encryption, since the scheme is a hybridization of speech information hiding and encryption, we evaluated the proposed scheme with respect to information hiding and encryption. We also evaluated the robustness of the entire system.

5.4.1 Scheme used for building secret and secure channels evaluation

Our proposed scheme is based on adding secured watermarks to the host speech signal for building secret and secure channels. The main objective is to achieve imperceptible hidden information in a watermarked signal. The attackers will not perceive the hidden information in the contents. Since the watermark bits were transformed before they were hidden into speech signals, the key is needed to discover the original watermark.

Three measurements were used to assess the imperceptibility of the watermark signal in a watermarked signal: the *log-spectral distance* (LSD), the *signal-to-distortion ratio* (SDR), and the *Perceptual Evaluation of Speech Quality* (PESQ). The LSD is the distance between the spectrum of the host speech and that of the watermarked signal (in dB). The BER of extract bits will be used to show the secure of the transformation. SDR is the power ratio between the signal and the distortion (in dB). The PESQ represents the sound-quality degradation of the watermarked signal compared with that of the host signal. The results principally model mean opinion scores (MOS) ranging from 1 (poor) to 5 (excellent).

The criteria for acceptable imperceptibility is as follows [60]. The LSD should be less than 1 dB, the SDR should be greater than 25 dB, and the PESQ should be greater than 3. The proposed method was evaluated on these measures, the results of which are shown in Table 5.1. The proposed method satisfies all three measures, which indicates that even though the speech contents could be heard, the hidden information was imperceptible. Additionally, the system's imperceptible properties are well-performing as in the pure SSA-based information hiding method of [60], and [46]. The BER

Table 5.1: Comparison of imperceptible properties between proposed and other methods

Method	LSD (dB)	SDR (dB)	PESQ
Parameterized SSA-based method [60]	0.65	31.58	3.70
SSA-based method [46]	0.69	30.96	3.64
Proposed method	0.65	31.56	3.70

without the correct key is higher than 50% that refer that the channel is secured.

5.4.2 Scheme deployed encryption evaluation

The correlation coefficient and signal-to-noise ratio (SNR) were measured to evaluate encryption and decryption. The correlation coefficient measures the linear relationship between the original speech, the encryption speech, and the decrypted speech, while SNR measures the noise content in the encrypted speech signal. The correlation coefficient between the original signal and the decrypted signal should be close to 1, which indicates no difference between the two signals, and the SNR should be high. On the other hand, the correlation coefficient between the original signal and the encrypted signal should be close to 0, which indicates the difference between the two, and the SNR should be small. Note that the original speech to be encrypted in this proposed method is the watermarked signal. Table 5.2 shows that the encryption and decryption performance of the proposed method was comparable to that of previously developed speech encryption methods [79] and [80].

Robustness of proposed scheme

The robustness of the proposed scheme was evaluated by the sensitivity of the encryption algorithm to changing one or multiple keys and the watermark-extraction precision of the information hiding. The following were measured to assess the sensitivity to key changes: the number of sample change rates (NSCR), the correlation coefficient, and the bit error rate (BER). The NSCR is defined by

$$\text{NSCR} = \frac{1}{L}\sum_{i=1}^{L} D_i, \tag{5.4}$$

Table 5.2: Comparison of correlation coefficient and SNR (in dB) between original speech (ori), encrypted speech (enc), and decrypted speech (dec) for the proposed method and other encryption methods. Note that NA is not applicable data

.

Method	Corr-coef (ori,enc)	SNR (ori,enc)	Corr-coef (ori,dec)	SNR (ori,dec)
Chaotic shift keying method [79]	0.04	NA	0.99	123.57
FFT with chaotic method [80]	0.02	NA	0.99	33.52
Proposed method	0.10	-2.52	0.99	31.74

where L corresponds to the length of the speech signal, and D_i is determined according to the rule

$$D_i = \begin{cases} 1, & \text{if } A_i \neq A_i{}', \\ 0, & \text{otherwise }, \end{cases} \tag{5.5}$$

where A_i and A'_i are are the amplitudes of the original speech and those of the encrypted speech, respectively.

The BER is defined as

$$\text{BER} = \frac{1}{M}\sum_{j=1}^{M} w(j) \oplus \hat{w}^*(j), \tag{5.6}$$

where $w(j)$ and $\hat{w}^*(j)$ are the embedded-watermark bits and the extracted-watermark bits, respectively.

BER was also used to represent the precision of watermark extraction in the proposed method. The NSCR and the correlation coefficient show the degree of variation between two encrypted speech signals when the keys are modified, and BER indicates the extraction precision when the keys were changed. Table 5.3 shows the measurements obtained when detecting the encrypted watermarked signal with a different key series (including the true key and wrong keys). If the true keys were applied, the ideal values for NSCR, correlation coefficient, and BER are 0%, 1, and 0%, respectively. The experiment results show that with the true keys, these three measurements are almost perfect values. The key changes slightly from the true keys to demonstrate the value when the wrong key was applied. The NCSR

Table 5.3: Key sensitivity and BER

.

Keys ($K_{A0}, K_{A1}, K_{B0}, K_{B1}$)	Corr-coef (range [0,1])	NSCR (%)	BER (%)
True key (5, 7, 11, 19)	1	0	0.13
True key (6, 6, 12, 18)	1	0	0.07
Wrong key(5, 6, 11, 18)	0.0072	99.29	50.72
Wrong key (6, 7, 11, 20)	0.0068	99.28	49.51
Wrong key (6, 9, 12, 19)	0.0069	99.27	50.15
Wrong key (6, 7, 12, 19)	0.0073	99.33	50.12

demonstrates that the two decrypted speech signals with slightly different keys hold different samples with near 100%, and the correlation coefficient is close to zero. The BER with the wrong keys is as high as 50%, while a BER with a true key is less than 1%. The measurement values indicate that the hidden information is secured.

5.5 Summary

In summary, there are two different applications in this chapter. The first application provides the secret and secure channel, while the second application provides the accessing data at the different levels.

The first application, SSA-based AIH is used to build the secret channel, and the Arnold transformation is performed on the watermark signal to make it secured. The hidden information was imperceptible so that it is secret because the listeners do not ware its existence. The BER without the correct key is higher than 50%, which means the channel is secured.

The second application, Arnold transformation was performed on watermark signals to create secured watermarks, which were then embedded into host speech using SSA-based information hiding, producing a watermarked signal. The watermarked signal was encrypted before being sent through the communication channel. The experimental results showed considerable differences between the correlation coefficient and SNR of the watermarked signal and those of the encrypted watermarked signal. The key sensitivity indicated that only authorized persons with the watermarked encryption key could access the speech contents. The imperceptible watermark and

significant difference of BER with and without a watermark key indicated that access to the hidden information was limited. This hybridization system increased speech security and limited accessibility to the data at varying levels.

Chapter 6

Conclusion

This chapter concludes this research work and highlights its contributions to the auditory information hiding research field and other research fields. However, no work will be complete at once, and this work can still be improved. Therefore, we show the possible way to improve our work in the future as well.

6.1 Summary

In this study, we propose to solve the security problem in the speech signal by using the information hiding method. The first security is to protect the genuineness of the speech signal. The hidden information is embedded in to host signal. If attackers modify or tampered with the speech signals, the hidden information will reflect the change to check for tampering. The second security is to protect the secret communication on the speech signal. In this study, we use our proposed information hiding method to build a secret and secured channel. Our core structure of SSA-base AIH and CNN-based parameter estimation are deployed in both scenarios: first, to detect tampering, and second, to build the secret and secured channel.

Since SSA is used to analyze and investigate the characteristic of a speech signal and the core structure of SSA-based AIH and CNN-based parameter estimation are deployed in both scenarios, we will summarize the basic facts that we find out about SSA to understand the concept of implementation on our AIH framework.

1. Singular value of speech signals is less sensitive to many signal processing attacks, i.e., its values are not easily changed. Thus, we hypothesis that if we hide information into singular value, hidden information should be maintained.

2. Singular spectrum is naturally convex. We hide the information by making concave on singular spectrum and we extract by checking the concave or convex on the modification part.

3. Singular values are sorted in descending order, and the lower-order singular values have more contributed to the signal than that of the higher-order and sum of all singular value from speech signal delivered by SSA can be reconstructed the speech signal.

4. The difference between two adjacent singular values in the lower-order is more than that of the adjacent two in higher-order.

5. Modification in lower-order singular value is easy to detect since the difference between two adjacent singulars is high, but this modification affects more on sound quality (bad imperceptibility).

6. Modification in higher-order singular value is difficult to detect since the difference between two adjacent singulars is low, and this modification is a negligible effect on sound quality (good imperceptibility).

These facts are fundamentally used to build SSA-based AIH for both objective. The followings are sum up to show the unique and novel points of this work.

1. A novel embedding rule with the embedding strength concept. The novel embedding rule will consider the character of the singular spectrum and apply the embedding strength, which is normal distributed on a modified part. The embedding strength will take a number of modified singular values, the mean, and the variance of those values into account.

2. SSA-based AIH is for tampering detection that requires semi-fragility. Therefore the selected part to be modified is different from the original SSA-based AIH. The different parts of the singular spectrum to be modified give different robustness and fragility. We need to investigate that the part that satisfied semi-fragile property. Therefore we use DE optimizer to simulate the cost function to meet the requirement, and then DE-based parameter estimation suggests a part to be modified that maintains the semi-fragility.

3. The parameter estimation using CNN to overcome the computational time for parameter estimation. These CNNs offer the parameter that suitable for each speech element and keep balancing of AIH requirement.

4. SSA-based AIH deployed the transformation method to provide a secret and secure channel and provide scheme for accessing data at different level. The novel point is that SSA-based AIH analyze the matrix represent host signal and Arnold transformation perform on a matrix, this transformation is suitable to cooperate with SSA-based AIH.

The following is sum up regarding each sub-goals to reach the ultimate goal.

1. SSA-based AIH applied on speech signal can keep the SSA-based technique advantage as it has done on an audio signal. Therefore, SSA-based AIH can be applied for both audio and speech signals. SSA-based AIH can be considered an advantage over the method that can only apply to audio or speech signals.

2. In order to detect tampering, the SSA-based AIH needs to achieve semi-fragility property. We found that our scheme is not robust to MP4 and also not fragile to echo adding attacks. Consequently, we cannot say that our proposed method successful in tampering detection. The FAR of our tampering detection system is 12.5 %, and this FAR is caused by our system cannot detect echo-adding attacks. The FRR is 5 %, and this FRR is caused by our tampering system judge MP4 operation as an attack. Therefore, the scheme should be more robust to MP4 while the scheme should be more fragile to echo-adding to improve tampering detection effectiveness.

3. Parameter obtained by the DE-based method gives good performance and reasonable to use for CNN training.

4. CNN-based parameter estimation can reduce the computational time and keep the requirement balancing as the original SSA-based AIH.

5. SSA-based AIH scheme can cooperate transformation to provide the secret channel and secret channel on the speech signal.

6.2 Contribution

The proposed information hiding scheme is mainly focused on tampering detection of speech signals. The contribution can be made to society by increasing the security of authentication systems by eliminating the tampered speech before feeding to authentication systems. This scheme can also check the originality of speech signals, and the scheme can be applied to faked

information detection in social networks. Moreover, it contributes to science as a novel analysis tool in digital forensics and checks the originality and integrity of speech recording in the court. The application we proposed in the last chapter is one example of our core structure providing the secret and secure channel on speech signal communication.

6.3 Future Work

The SSA-based AIH proposed in this work still have rooms for futher enhancements as follow.

1. The core structure that applied CNN-based parameter estimation [73] has shown a good sign of robustness scheme against speech coding. Comparison with our previous work [60], robustness against G.726 speech coding is significantly increased (from 21.07 % to be 12.12 %) while the robustness against G.711 is maintained. In addition, we succeeded in reducing the computational time of parameter estimation. Since we deployed DE to generate a dataset for training our CNNs, the scheme's effectiveness is correlated with DE performance, and DE performance depends on its cost function. Therefore, if we want our scheme to robust against we must set up a good cost function to ensure the robustness against speech coding. Our DE optimizer considers only two types of speech coding in this core structure, but the scheme shows a good sign of robustness improving. We assume the better cost function will offer better robustness. Moreover, the semi-fragility that the scheme should be more robust to MP4 and more fragile to echo-adding can be simulated using DE optimizer.

2. The core structure that applied CNN-based parameter estimation [73] has a limitation on embedding capacity. The CNNs parameter is trained with a high-quality dataset generated from the DE optimizer, and our DE optimizer is simulated with a fixed frame length at 400 samples per one frame. We hypothesis that frame length affects the estimated parameter because if the frame length changed, the number of possible modified singular would also change, then the accuracy of parameter estimation is concerned. Therefore, the problem of different frame lengths or different embedding capacities should be tackled in future work.

3. In CNN parameter estimation, there are three parameters, γ, μ, and σ, to be estimated, and their value is quite small compared to parameter

γ. Accordingly, we implement two CNNs, one for μ and σ and the other for γ. We can reduce the CNNs to be only one CNN by using the weight function. However, its weight function must be design properly to balance robustness and sound quality because parameters μ and σ relate to the embedding strength α_i, which contribute to the robustness of the proposed scheme. In contrast, parameter γ directly defines the number of modified singular values and contributes more to the sound quality. Therefore, the weight function can help to reduce the CNNs to be only one CNN. Moreover, now the CNN was used only for parameter estimation of information hiding schemes, which is used for tampering detection. If the CNN can co-operate the scheme on tampering detection function, the scheme's performance for tampering detection can be improved.

4. In chapter 5, we applied our SSA-based AIH with the transformation method to provide a secret and secure channel on the speech signal. Arnold transformation is performed to provide secured watermarks. The secured watermarks are embedded into the host signal to obtain a watermarked signal. Consequently, the channel is secret because hidden information does not attract the listener, and transformation makes the hidden information secure. Only the authorized person with the key can access the hidden information. This SSA-based method AIH scheme succeeds in building a secret and secure channel to protect the secret communication. Moreover, by performing a transformation on watermarked signals, the scheme can limit accessing data at varying levels. However, the transformation method we used in this proposed application deployed Arnold transformation, and Arnold transformation is cyclic and iteration. Arnold's transformation has a weak point: it cannot provide high security against brute-force attacks because it has the limitation of the key searching space. Therefore, there are wide encryption algorithms that can be deployed for better security.

Bibliography

[1] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: a survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 746–789, 2019.

[2] M. Buchholz, I. M. Müller, and U. Ferm, "Text messaging with pictures and speech synthesis for adolescents and adults with cognitive and communicative disabilities–professionals' views about user satisfaction and participation," *Technology and Disability*, vol. 25, no. 2, pp. 87–98, 2013.

[3] A. Alsaif, N. Albadrani, A. Alamro, and R. Alsaif, "Towards intelligent arabic text-to-speech application for disabled people," in *2017 International Conference on Informatics, Health & Technology (ICIHT)*. IEEE, 2017, pp. 1–6.

[4] B. Busatlic, N. Dogru, I. Lera, and E. Sukic, "Smart homes with voice activated systems for disabled people," *TEM Journal*, vol. 6, no. 1, p. 103, 2017.

[5] G. Roma, O. Green, and P. A. Tremblay, "Audio morphing using matrix decomposition and optimal transport," in *23rd International Conference on Digital Audio Effects*, 2020, pp. 147–154.

[6] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[7] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[8] D. Ormerod, "Sounding out expert voice identification," *Expert Evidence and Scientific Proof in Criminal Trials*, 2017.

[9] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2017, pp. 241–250.

[10] Z. Qi and C. Qi, "Evaluate the managerial effectiveness of artificial intelligence-case studies from hsbc and iqiyi," in *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, 2019, pp. 542–547.

[11] L. Valentine, "Voice recognition frees hands; boosts customer service," *American Bankers Association. ABA Banking Journal*, vol. 94, no. 4, p. 51, 2002.

[12] O. G. Abood and S. K. Guirguis, "A survey on cryptography algorithms," *International Journal of Scientific and Research Publications*, vol. 8, no. 7, pp. 495–516, 2018.

[13] V. M. Rumata and A. S. Sastrosubroto, "The indonesian law enforcement challenges over encrypted global social networking platforms," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. IEEE, 2018, pp. 199–203.

[14] M. Mann, A. Daly, and A. Molnar, "Regulatory arbitrage and transnational surveillance: Australia's extraterritorial assistance to access encrypted communications," *Internet Policy Review*, vol. 9, no. 3, pp. 1–20, 2020.

[15] R. Thanki, K. Borisagar, and S. Borra, *Advance compression and watermarking technique for speech signals*. Springer, 2018.

[16] R. C.-W. Phan, Y.-Y. Low, K. Wong, and K. Minemura, "Strengthening speech content authentication against tampering," *Speech Communication*, 2021.

[17] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Transactions on information theory*, vol. 49, no. 3, pp. 563–593, 2003.

[18] Z. Wu, *Information Hiding in Speech Signals for Secure Communication*. Syngress, 2014.

[19] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[20] R. Namikawa and M. Unoki, "Non-blind speech watermarking method based on spread-spectrum using linear prediction residue," *IEICE Transactions on Information and Systems*, vol. 103, no. 1, pp. 63–66, 2020.

[21] S. Wang, W. Yuan, J. Wang, and M. Unoki, "Speech watermarking based on robust principal component analysis and formant manipulations," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2082–2086.

[22] C. H. Yeh and C. Kuo, "Digital watermarking through quasi m-arrays," in *1999 IEEE Workshop on Signal Processing Systems. SiPS 99. Design and Implementation (Cat. No. 99TH8461)*. IEEE, 1999, pp. 456–461.

[23] S. Wang and M. Unoki, "Watermarking method for speech signals based on modifications to lsfs," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2013, pp. 283–286.

[24] J. Karnjana, P. Aimmanee, M. Unoki, and C. Wutiwiwatchai, "An audio watermarking scheme based on automatic parameterized singular-spectrum analysis using differential evolution," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 543–551.

[25] L. Lamarche, Y. Liu, and J. Zhao, "Flaw in svd-based watermarking," in *2006 Canadian Conference on Electrical and Computer Engineering*. IEEE, 2006, pp. 2082–2085.

[26] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding-a survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[27] E. Erçelebi and L. Batakçı, "Audio watermarking scheme based on embedding strategy in low frequency components with a binary image," *Digital Signal Processing*, vol. 19, no. 2, pp. 265–277, 2009.

[28] W.-N. Lie and L.-C. Chang, "Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification," *IEEE transactions on multimedia*, vol. 8, no. 1, pp. 46–59, 2006.

[29] M. Fallahpour and D. Megias, "High capacity audio watermarking using fft amplitude interpolation," *IEICE Electronics Express*, vol. 6, no. 14, pp. 1057–1063, 2009.

[30] E. Erçelebi and A. Subaşı, "Robust multi bit and high quality audio watermarking using pseudo-random sequences," *Computers & Electrical Engineering*, vol. 31, no. 8, pp. 525–536, 2005.

[31] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking—a comprehensive review," *Signal processing*, vol. 128, pp. 222–242, 2016.

[32] D.-Y. Huang and T. Y. Yeo, "Robust and inaudible multi-echo audio watermarking," in *Pacific-Rim Conference on Multimedia*. Springer, 2002, pp. 615–622.

[33] S. N. Neyman, I. N. P. Pradnyana, and B. Sitohang, "A new copyright protection for vector map using fft-based watermarking," *Telkomnika*, vol. 12, no. 2, p. 367, 2014.

[34] A. Merrad and S. Saadi, "Blind speech watermarking using hybrid scheme based on dwt/dct and sub-sampling," *Multimedia Tools and Applications*, vol. 77, no. 20, pp. 27 589–27 615, 2018.

[35] B. Y. Lei, Y. Soon, and Z. Li, "Blind and robust audio watermarking scheme based on svd–dct," *Signal Processing*, vol. 91, no. 8, pp. 1973–1984, 2011.

[36] M. A. Nematollahi, S. Al-Haddad, and F. Zarafshan, "Blind digital speech watermarking based on eigen-value quantization in dwt," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 58–67, 2015.

[37] L. Cui, S.-X. Wang, and T. Sun, "The application of wavelet analysis and audio compression technology in digital audio watermarking," in *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, vol. 2. IEEE, 2003, pp. 1533–1537.

[38] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE transactions on signal processing*, vol. 51, no. 4, pp. 1020–1033, 2003.

[39] M. Unoki and R. Miyauchi, "Method of digital-audio watermarking based on cochlear delay characteristics," in *Multimedia Information Hiding Technologies and Methodologies for Controlling Data*. IGI Global, 2013, pp. 42–70.

[40] R. Nishimura, "Audio watermarking using spatial masking and ambisonics," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 9, pp. 2461–2469, 2012.

[41] X. Dong, M. F. Bocko, and Z. Ignjatovic, "Data hiding via phase manipulation of audio signals," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 2004, pp. V–377.

[42] R. Ansari, H. Malik, and A. Khokhar, "Data-hiding in audio using frequency-selective phase alteration," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 2004, pp. V–389.

[43] D. S. Broomhead, G. P. King *et al.*, "On the qualitative analysis of experimental dynamical systems," *Nonlinear phenomena and chaos*, vol. 113, p. 114, 1986.

[44] N. Alharbi and H. Hassani, "A new approach for selecting the number of the eigenvalues in singular spectrum analysis," *Journal of the Franklin Institute*, vol. 353, no. 1, pp. 1–16, 2016.

[45] M. Unoki and R. Miyauchi, "Detection of tampering in speech signals with inaudible watermarking technique," in *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2012, pp. 118–121.

[46] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "Tampering detection in speech signals by semi-fragile watermarking based on singular-spectrum analysis," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Springer, 2017, pp. 131–140.

[47] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[48] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[49] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 309–312.

[50] P. L. D. Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[51] C.-P. Wu and C.-C. J. Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–3305.

[52] Z. Liu and H. Wang, "A novel speech content authentication algorithm based on bessel–fourier moments," *Digital Signal Processing*, vol. 24, pp. 197–208, 2014.

[53] S. Wang, R. Miyauchi, M. Unoki, and N. S. Kim, "Tampering detection scheme for speech signals using formant enhancement based watermarking," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 6, pp. 1264–1283, 2015.

[54] S. Wang, W. Yuan, J. Wang, and M. Unoki, "Speech watermarking based on source-filter model of speech production." *J. Inf. Hiding Multim. Signal Process.*, vol. 10, no. 4, pp. 517–534, 2019.

[55] C. O. Mawalim, S. Wang, and M. Unoki, "Speech information hiding by modification of lsf quantization index in celp codec," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 1321–1330.

[56] S. Wang, W. Yuan, J. Wang, and M. Unoki, "Detection of speech tampering using sparse representations and spectral manipulations based information hiding," *Speech Communication*, vol. 112, pp. 1–14, 2019.

[57] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "An audio watermarking scheme based on singular-spectrum analysis," in *International Workshop on Digital Watermarking*. Springer, 2014, pp. 145–159.

[58] M. Moonen and B. De Moor, *SVD and Signal Processing, III: Algorithms, Architectures and Applications*. Elsevier, 1995.

[59] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "Singular-spectrum analysis for digital audio watermarking with automatic

parameterization and parameter estimation," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 8, pp. 2109–2120, 2016.

[60] J. Karnjana, K. Galajit, P. Aimmanee, C. Wutiwiwatchai, and M. Unoki, "Speech watermarking scheme based on singular-spectrum analysis for tampering detection and identification," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 193–202.

[61] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "Ssa-based audio-information-hiding scheme with psychoacoustic model," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–10.

[62] R. Burimas, T. Kumpuak, M. Intarauksorn, K. Galajit, P. Aimmanee, and J. Karnjana, "Framework for hiding information in audio sub-signals by using singular spectrum analysis with psychoacoustic model," in *Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing*, 2019, pp. 67–74.

[63] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "Audio watermarking scheme based on singular spectrum analysis and psychoacoustic model with self-synchronization," *Journal of Electrical and Computer Engineering*, vol. 2016, 2016.

[64] K. Galajit, J. Karnjana, P. Aimmanee, and M. Unoki, "Digital audio watermarking method based on singular spectrum analysis with automatic parameter estimation using a convolutional neural network," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Springer, 2018, pp. 63–73.

[65] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[66] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[67] K. Iwamura, M. Kawamura, M. Kuribayashi, M. Iwata, H. Kang, S. Gohshi, and A. Nishimura, "Information hiding and its criteria for evaluation," *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 1, pp. 2–12, 2017.

[68] P. Bassia, I. Pitas, and N. Nikolaidis, "Robust audio watermarking in the time domain," *IEEE Transactions on multimedia*, vol. 3, no. 2, pp. 232–241, 2001.

[69] S. Wang, M. Unoki, and N. S. Kim, "Formant enhancement based speech watermarking for tampering detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[70] "ATR Japanese Speech Database (set B) ATR-Promotions, Inc., Japan," http://www.atr-p.com/products/sdb.html#DIGI, accessed: 2021-03-13.

[71] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[72] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[73] K. Galajit, J. Karnjana, M. Unoki, and P. Aimmanee, "Semi-fragile speech watermarking based on singular-spectrum analysis with cnn-based parameter estimation for tampering detection," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[74] M. M. Rahman, T. K. Saha, and M. A.-A. Bhuiyan, "Implementation of rsa algorithm for speech data encryption and decryption," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 12, no. 3, p. 74, 2012.

[75] N. Radha and M. Venkatesulu, "A chaotic block cipher for real-time multimedia," *Journal of Computer Science*, vol. 8, no. 6, p. 994, 2012.

[76] Y. Lin and W. H. Abdulla, "A secure and robust audio watermarking scheme using multiple scrambling and adaptive synchronization," in *2007 6th International Conference on Information, Communications & Signal Processing*. IEEE, 2007, pp. 1–5.

[77] R. Shelke and M. Nemade, "Audio encryption algorithm using modified elliptical curve cryptography and arnold transform for audio

watermarking," in *2018 3rd International Conference for Convergence in Technology (I2CT)*. IEEE, 2018, pp. 1–4.

[78] Q. Wu and M. Wu, "Adaptive and blind audio watermarking algorithm based on chaotic encryption in hybrid domain," *Symmetry*, vol. 10, no. 7, p. 284, 2018.

[79] P. Sathiyamurthi and S. Ramakrishnan, "Speech encryption using chaotic shift keying for secured speech communication," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, pp. 1–11, 2017.

[80] P. Sathiyamurthi and S. Ramakrishnan, "Speech encryption algorithm using fft and 3d-lorenz–logistic chaotic map," *Multimedia Tools and Applications*, pp. 1–19, 2020.

PUBLICATIONS

Main Publications

International Journals

1. Galajit, K., Karnjana, J., Unoki, M., and Aimmanee, P., "Semi-fragile speech watermarking based on singular-spectrum analysis with CNN-based parameter estimation for tampering detection," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

Books Chapter

1. Galajit, K., Karnjana, J., Aimmanee, P., and Unoki, M., "Digital audio watermarking method based on singular spectrum analysis with automatic parameter estimation using a convolutional neural network," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Japan, pp. 63-73, 2018. Springer, Cham.

International Conference Proceedings

1. Karnjana, J., Galajit, K., Aimmanee, P., Wutiwiwatchai, C., and Unoki, M., "Speech watermarking scheme based on singular-spectrum analysis for tampering detection and identification," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, KL Malaysia, pp. 193-202, 2017, IEEE.

2. Galajit, K., Karnjana, J., Aimmanee, P., and Unoki, M., "Digital audio watermarking method based on singular spectrum analysis with automatic parameter estimation using a convolutional neural network," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Japan, pp. 63-73, 2018, Springer, Cham.

3. Galajit, K., Karnjana, J., Unoki, M., Intarauksorn, M., and Aimmanee, P., "Speech watermarking technique based on singular spectrum analysis and automatic parameter estimation using differential evolution for tampering detection," in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, BKK Thailand, pp. 1-6, 2018, IEEE.

4. Burimas, R., Kumpuak, T., Intarauksorn, M.,Galajit, K., Aimmanee, P., and Karnjana, J., "Framework for Hiding Information in Audio Sub-signals by Using Singular Spectrum Analysis with Psychoacoustic Model," in *Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing*,pp. 67-74, 2019.

5. Mawalim, C. O., Galajit, K., Karnjana, J., and Unoki, M., "X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System," in *SProc. Interspeech 2020, pp. 1703-1707, 2020.*

Domestic Conference

1. Galajit, K., Karnjana, J., Aimmanee, P., and Unoki, M., "Study on Digital Audio Watermarking Method Based on Singular Spectrum Analysis with Automatic Parameter Estimation Using a Convolutional Neural Network," in *Technical Committee on Multimedia Information Hiding and Enrichment (EMM)*, Japan, 2019.

2. Galajit, K., Karnjana, J., Aimmanee, P., and Unoki, M., "Study on singular spectrum analysis-based speech watermarking technique with parameter estimation using differential evolution," in *2019 Spring Meeting Acoustic Society of Japan(ASJ)*, Japan, 2019.

3. Galajit, K., Karnjana, J., and Unoki, M., "Audio Information Hiding in Sub-signals by deploying Singular Spectrum Analysis and Psychoacoustic Model," in *5th Technical Committee on Multimedia Information Hiding and Enrichment (EMM)*, Japan, 2021.

Awards Information

(a) **Best Paper Award in IIH-MSP 2018.**
Galajit, K., Karnjana, J., Aimmanee, P., and Unoki, M., "Digital

audio watermarking method based on singular spectrum analysis with automatic parameter estimation using a convolutional neural network," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Japan, pp. 63-73, 2018, Springer, Cham.

(b) **Best Paper Award in iSAI-NLP 2018.**
Galajit, K., Karnjana, J., Unoki, M., Intarauksorn, M., and Aimmanee, P., "Speech watermarking technique based on singular spectrum analysis and automatic parameter estimation using differential evolution for tampering detection," in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, BKK Thailand, pp. 1-6, 2018, IEEE.

(c) **Best Presentation Award in iSAI-NLP 2018.**
Galajit, K., Karnjana, J., Unoki, M., Intarauksorn, M., and Aimmanee, P., "Speech watermarking technique based on singular spectrum analysis and automatic parameter estimation using differential evolution for tampering detection," in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, BKK Thailand, pp. 1-6, 2018, IEEE.

(d) **Excellent Student Presentation Award in the 5th EMM Conference 2020FY.**
Galajit, K., Karnjana, J., and Unoki, M., "Audio Information Hiding in Sub-signals by deploying Singular Spectrum Analysis and Psychoacoustic Model," in *5th Technical Committee on Multimedia Information Hiding and Enrichment (EMM)*, Japan, 2021.